刘颖 著

基于机器学习的遥感影像分类方法研究

清华大学出版社

刘颖 著

基于机器学习的遥感影像分类方法研究

清华大学出版社
北 京

内 容 简 介

机器学习是人工智能的一个重要领域，源自于统计模型拟合。机器学习通过采用推理及样本学习等方式从数据中获得相应的理论，尤其适合解决“噪声”模式及大规模数据集等问题。本书是作者几年来科研成果的总结。全书共 7 章，围绕遥感图像分类这一主线，深入研究监督学习、半监督学习、集成学习三大主流机器学习算法，构建完整的遥感图像分类体系。在理论研究的基础上，结合实例，详细介绍了改进机器学习算法及其在遥感分类处理中的应用情况。

本书内容充实、结构清晰、实例丰富，适合从事计算机及相关学科的师生，以及相关科研院所的科研人员阅读。

图书在版编目(CIP)数据

基于机器学习的遥感影像分类方法研究 / 刘颖 著. —北京：清华大学出版社，2014

ISBN 978-7-302-35991-3

Ⅰ. ①基… Ⅱ. ①刘… Ⅲ. ①机器学习—应用—遥感图像—分类—研究 Ⅳ. ①TP75②TP181

中国版本图书馆 CIP 数据核字(2014)第 065957 号

责任编辑：王桑娉　易银荣
封面设计：牛艳敏
版式设计：方加青
责任校对：曹　阳
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：169mm×230mm　　印　张：10.25　　字　数：114 千字
版　　次：2014 年 5 月第 1 版　　印　次：2014 年 5 月第 1 次印刷
定　　价：58.00 元

产品编号：058438-01

前　言

长期以来，土地覆盖变化的研究一直是全球环境研究的热点，无论从社会经济角度还是从生态环境角度均具有重要的意义。为了全面掌握土地覆盖变化信息，迫切需要使用切实有效的方法实现土地覆盖宏观、动态、大尺度的制图与监测，遥感技术的迅猛发展为这一需求提供了可能。然而，目前遥感信息处理和分类的水平大大滞后于遥感影像获取技术的发展。因此，研究新理论、新方法以提高遥感信息的处理能力具有十分重要的意义和应用前景。

支持向量机(Support Vector Machines，SVM)是近年来机器学习与模式识别领域新的研究焦点，它具有结构简单、适应性强、全局最优等特点，能较好地解决高维特征、非线性，过学习与不确定性等问题，广泛地应用于土地覆盖遥感分类。尽管SVM在遥感信息获取中取得了很好的效果，但仍存在有待改进和完善之处，主要表现在以下两方面：①参数选择的问题，即不准确的分类参数常常影响分类器的分类精度；②样本不足且代表性不好的问题，即当训练样本集远远小于测试样本集时，即便SVM具有较强的泛化性，也难以给出令人满意的结果。围绕这些问题，本书开展了如下工作：

1. 选择图们江下游，中、朝、俄交界处作为研究对象。以行列号115-30一景、近20年的6幅不同时相的Landsat ETM/TM影像作为研究材料。分别讨论本书所采用的影像合成方式、特征采集方法、土地覆盖分类依据，以及特征选取方法，为进一步研究分类方法提供必要的数据材料。

2. 针对SVM分类过程中核函数选择及参数设置不准确的缺点，提出一种基于自适应变异粒子群优化SVM参数模型(Adaptive Mutation Particle Swarm Optimization SVM，AMPSO-SVM)。AMPSO在运行过程中根据群体适应度方差以及最优解的大小来确定当前最佳粒子的变异概率。与传统粒子群(Particle Swarm Optimization，PSO)优化SVM参数模型(PSO-SVM)相比，AMPSO-SVM能够快速摆脱局部搜索的束缚，提高全局搜索的性能，克服早熟收敛造成分类参数寻找不准确的缺点，同时保持了种群的多样性。最后应用该模型进行多光谱遥感影像的土地覆盖分类实验，并与SVM分类方法、PSO-SVM分类方法进行对比。分类精度从传统PSO-SVM的91.50%提高到93.59%，Kappa系数由0.890 3提高为0.917 5。c和γ的取值得到的分类结果明显优于SVM的手工设置值100和0.143所得到的结果(分类精度87.07%，Kappa系数0.837 2)。结果表明，AMPSO-SVM模型有效地提高了遥感影像的分类精度。

3. 提出了一个新的自训练半监督支持向量机方法(PS3VM)。自训练半监督算法最大的弊端在于“错误累积”现象，即在学习过程

中，一旦某个分类出错，将导致这个错误被继续学习与加强。为了克服这一现象，本书在自训练半监督SVM(S3VM)的基础上引入两个算法：①从分类器的构造角度，利用自适应变异粒子群算法对SVM参数优化，以提高单个分类器的分类精度；②在未标记样本的标注阶段，采用Gustafson-Kessel模糊聚类算法(GKclust)将最接近样本的有效无标签样本作为标注对象，以控制错误信息的输入。为了测试所提模型的有效性，分别针对遥感的数字化集合和影像集合进行分类实验，并与AMPSO-SVM(简称PSVM)监督分类方法、未改进自训练S3VM方法进行对比实验，由PS3VM产生的分类精度(95.10%)分别比S3VM(93.06%)高出2.04%；比PSVM(90.81%)高出4.29%。实验结果一方面说明了已标记样本和未标记样本的用量比例必须满足一定的阈值要求(1:3)，才能产生最小的泛化误差；另一方面证实了利用所提出学习框架能够获得较好的分类精度。

4. 对于样本不足且代表不好而造成的小样本问题，学者们普遍采用半监督学习和集成学习两种范式对SVM进行改进。然而，集成学习与半监督学习之间存在许多互补性，且二者的混合范式(即半监督集成)可以更大程度地改进学习系统的泛化能力。因此，本文设计了一种新的半监督集成方案(EPS3VM)，PS3VM半监督方法利用未标记数据有效地应对训练样本不足的同时也产生若干性能差异的个体分类器，将这些个体分类器采用加权集成策略进一步提高分类模型的泛化能力。为了测试其性能，应用该模型进行多光谱

遥感影像的土地覆盖分类实验，并与其相关算法进行对比。分类精度从92.16%(PS3VM)提高到96.88%，Kappa系数由0.901 0提高为0.960 6。结果表明，EPS3VM克服传统SVM参数选择不准确的同时有效地应对了小样本问题，分类性能更优。

本书是在吉林财经大学资助下，国家自然科学基金项目(61202306)、吉林省科技厅项目(20130522177JH，201215119，20100507)、吉林省教育厅十二五重点规划项目(2012185)、吉林省高校新世纪优秀人才支持计划、吉林财经大学青年学俊等项目的支持下完成的。值此专著完成之际，诚挚地感谢吉林财经大学的资金支持，感谢中国科学院东北地理与农业生态研究所张柏教授、吉林财经大学管理科学与信息工程学院王丽敏教授及长春工业大学韩旭明副教授的热情帮助和指点。

由于作者水平有限，加之机器学习领域研究领域纵深宽广，书中难免有考虑不周之处，诚请广大读者批评指正。

刘　颖

2014年1月于长春

目　录

第1章

绪　论

1.1 基本概念

1.1.1 土地覆盖

土地覆盖变化是全球环境变化研究的热点和前沿，是国际地圈—生物圈计划(IGBP)的核心研究内容之一，无论从社会经济角度还是从生态环境角度均具有重要的意义。土地覆盖真实地反映了地表覆盖情况，它并不是单一的植被及土地类型，而是土地利用类型及其所具有的一系列自然属性特征的综合体，包括与土地覆盖类型密切相关的生态环境要素，如植被所处的生态区域、地形与气候条件以及土地利用状况等[1]。

自20世纪80年代以来，随着臭氧层的破坏、全球变暖、土地退化、土地荒漠化的蔓延、酸雨频繁发生等全球性环境问题的不断涌现，人们逐渐认识到土地利用与土地覆盖变化是气候、生态过程、生物多样性、生物化学循环，乃至全球变化的主要原因[2]~[3]。具体表现在以下几个方面。

(1) 土地覆盖变化对人类资源利用的影响。由于土地覆盖变化使得水资源、耕地资源、草场资源、森林资源的存量发生变化，从而对人类社会产生影响，并对人类环境的安全造成威胁[4]~[5]。

(2) 土地覆盖变化对土壤的影响。土地覆盖的变化使得土壤养分失衡、营养元素衰竭、有机质含量变化、水分循环改变[6]~[8]。

(3) 土地覆盖变化对全球气候变化的影响。地表反照率、粗糙度、植被叶面积指数及植被覆盖度等变化造成降水量减少和温度增加[9] [10]，同样也引起局地区域气候变化，如城市热岛现象，城市酸

雨等[11]~[12]。

(4) 土地覆盖变化对生态系统的影响，集中于生态系统功能方面的物质循环和能量流动。如在温度、CO_2浓度及降水量等自然环境要素发生变化的情况下，生态系统的初级生产量会受到影响[13]~[14]。

(5) 土地覆盖变化对水文和地表、地下径流的影响。表现在其改变所造成的气候变化，汛期流量增加，枯水期流量减少，水中泥沙含量变化，径流极值变化，进而影响地下水的补给，产生多种水文效应[15]~[17]。

(6) 土地覆盖变化还可能加剧自然灾害。如森林的砍伐、植被覆盖度的降低使得土壤侵蚀面积增大，洪涝灾害加剧[18]~[19]。

因此，有效地管理和保护日益稀缺的土地资源，探索土地覆盖动态变化过程，解析变化的驱动因素，抑制全球环境恶劣变化等诸多问题一直是地学领域学者关注的热点，而这些问题的解决依赖于准确、实时地获取土地覆盖动态变化信息。

1.1.2　遥感技术

遥感RS(Remote Sensing)是20世纪60年代兴起并迅速发展起来的一门包括航天技术、计算机技术、图像处理等综合性的探测技术[20]。它是一种远离目标，在不与目标直接接触的情况下，通过某种平台上装载的传感器获取其特征信息，然后对其所获得的信息进行提取、判定、加工处理及应用分析[21]。遥感技术能大面积、重复地获取区域多波段、多时相信息，为大面积、实时、动态监测土地覆盖动态变化提供了可能，是调查、监测和分析最好的手段之一[22]。从此，人类对地球表层的理解进入一个崭新的阶段。然而，由于地球

系统的复杂性和开放性，遥感信息在进行地学空间分析和反演过程中具有模糊性和多解性的特点，即复杂性和不确定性[23]~[25]。表现为不同土地覆盖类型经常有相似的光谱特性(异物同谱现象)，以及相同土地覆盖类型由于地形、光照条件等的影响会具有不同的光谱特性(同物异谱现象)，这些增大了土地覆盖信息的提取难度。因此，在研究不同土地覆盖类型的光谱特征和空间分布特征的同时，进一步研究土地覆盖分类策略与方法，将有助于土地覆盖信息的有效获取。

1.1.3 机器学习

机器学习是人工智能的一个重要领域，源自于统计模型拟合。机器学习通过采用推理及样本学习等方式从数据中获得相应的理论，尤其适合解决“噪声”模式及大规模数据集等问题。在大样本、多向量及不确定数据分析工作中发挥着日益重要的作用[26]~[27]。

根据学习方式的不同，机器学习方法可分为以下几类：①监督学习(Supervised Learning)，利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程；②无监督学习(Unsupervised Learning)，是指所有样本的类别都是未知的，算法通过数据特征自动产生类别属性；③半监督学习(Semi-supervised Learning)，介于监督学习和无监督学习之间，所需样本既包括已知类别样本又包括未知类别样本，通过挖掘未知类别样本中所蕴涵的固有结构信息，来对已知类别样本可能因代表性不好而造成的拟合分类器有偏差的情况进行校正；④集成学习(Ensemble Learning)，综合多个同构或异构学习机对同一个问题进行学习进而提高分类器的泛化能力。

当前，许多机器学习技术被广泛地应用于遥感影像的分类，例如最大似然法、K-means算法、神经网络、决策树、支持向量机等[28]~[29]。其中，Vapnik提出的支持向量机(Support Vector Machines，SVM)监督分类技术是近年来模式识别与机器学习领域一个新的研究热点[30]~[31]，非常适合处理高维、复杂的小样本多维数据分类[32]~[33]，广泛应用于土地覆盖分类[34]~[36]、森林类型识别[37]~[38]、农业作物监测[39]、道路信息提取[40]、图像分割[41]等领域。尽管SVM在遥感信息提取中取得了很好的效果，但是仍然存在需要改进和完善之处。主要表现在以下两方面。

(1) 参数选择的问题：分类参数的选择没有特别好的办法，应用时不容易找到最优分类参数。

(2) 样本不足且代表不好的问题：当训练样本集远远小于测试样本集，即便SVM具有较强的泛化性，也难以给出令人满意的结果。

1.2 研究意义

本研究充分考虑遥感影像特点及SVM土地覆盖分类技术不足的问题，一方面利用智能优化算法，克服传统SVM参数选择不准确弊端；另一方面，将模糊聚类技术、半监督学习和集成学习理论引入SVM土地覆盖分类方法中，有效地解决小样本问题。其目的在于丰富土地覆盖遥感信息提取理论与方法，提高土地覆盖分类精度，为土地资源的合理保护、利用与监测提供更为有效的数据信息。

1.2.1 丰富土地覆盖遥感分类的理论与方法

综合现有的遥感影像分类研究工作，许多学者都默认所选择的样本对影像中地物类别具有较好的代表性，而将分类结果不理想归咎于分类算法的不适用。然而，在实际应用中，样本的选取不可避免地会存在易出错和经验知识有限等缺点，很难保证获得的训练样本能对分类类别有很好的代表性。针对遥感影像分类中人为选择样本代表性不好且样本数量不足等问题，本书提出将半监督学习、模糊聚类、集成学习与SVM分类方法相结合，减少了分类对准确先验知识或充足的训练样本的依赖程度，节省了大量样本获取所需的人力、物力和财力。

参数选择是任何机器学习方法都必须面对的重要而困难的问题。选择适合的SVM核函数及其设置准确参数对提高遥感影像分类精度非常重要。本选题所研究的利用智能优化算法对SVM的参数优化极大地改善了传统SVM分类的泛化能力。

1.2.2 为土地利用/覆盖的动态监测、保护和管理提供技术支持

通过有效的遥感分类方法准确提取土地覆盖信息，掌握土地利用/覆盖变化程度、类型和分布，为因土地覆盖变化而造成的土地退化、荒漠化，甚至全球变暖、酸雨等环境问题的预防和治理提供参考数据和科学依据，此研究具有重要的理论意义与实际应用价值。

1.2.3　一种新的自适应半监督支持向量机遥感分类模型的提出

著作提出一种新的半监督支持向量机分类模型，在自训练半监督算法的基础上，利用自适应变异粒子群算法对SVM参数优化，提高单个分类器的分类精度；在Self-training标注未标记样本阶段，采用Gustafson-Kessel模糊聚类算法将最接近样本的有效无标签样本作为标注对象，以控制错误信息的输入，实验结果表明其土地覆盖分类效果较好。

1.2.4　半监督学习思想和集成学习思想的融合

多年以来，半监督学习与集成学习的发展几乎是并行的，只有少数研究涉及二者的融合。本书充分分析集成学习与半监督学习之间的关系，提出将半监督分类器集成，构建半监督集成土地覆盖分类模型，进一步改进学习系统的泛化能力。

1.3　本书研究方法及结构安排

1.3.1　研究方法

1. 构建遥感影像数字集

本书以图们江下游，中、朝、俄交界处作为研究对象，选取行列号115-30遥感影像，包括近20年的6幅不同时相的Landsat TM/

ETM+影像，依据联合国粮农组织提出的土地覆盖分类体系和中国科学院资源环境数据库土地利用分类体系，从遥感制图角度和区域地理环境特点出发，建立了研究区土地覆盖遥感分类体系，利用简单随机采样方法进行样本提取，采用主成分分析(PCA)、NDVI植被指数等方法获取影像分类特征集合，如图1-1(a)部分所示。

2. SVM算法

SVM以统计学习VC维理论和结构风险最小化原理为原则，根据有限样本信息在模型的复杂性(即对特定训练样本的学习精度)和期望风险(即无错误地识别任意样本的能力)之间寻求最佳折中，以获得更好的泛化性能。与传统遥感分类方法相比，SVM技术具有维数不敏感、泛化能力强等优点，适合处理高维、小样本等问题，因而，成为近几年遥感分类领域中的一个非常活跃的研究热点，如图1-1(b)部分所示。

3. 自适应变异粒子群算法

自适应变异粒子群算法(Adaptive Mutation Particle Swarm Optimization，AMPSO)在运行过程中根据群体适应度方差以及当前最优解的大小来确定当前最佳粒子的变异概率，当算法陷入局部极值时，根据变异算子使得算法能够及时跳出局部极值进而获得全局最优解。其目的是快速摆脱局部搜索的束缚，提高全局搜索的性能，进而克服早熟收敛造成分类参数寻找不准确的缺点，如图1-1(b)部分所示。

4. Gustafson-kessel聚类算法

模糊性是遥感图像所固有的特性，由于不确定性和混合像元的存在，多光谱遥感图像散点图在特征空间中是趋于椭球体分布的，

而基于欧式距离的Fuzzy c-means算法更适合于对球体分布的数据进行聚类分析。Gustafson-Kessel(GKclust)是距离自适应动态聚类算法(Adaptive Distance Dynamic Clustering Algorithm)的模糊推广，可以有效地搜索超椭球、平面或线型的数据类，在遥感数据聚类表现效果更优，如图1-1(c)部分所示。

5. 自训练半监督学习算法

自训练(Self-training)是半监督学习比较简单且常用的方法。自训练方法首先用有标签样本训练一个弱分类器，然后用此弱分类器对所有无标签样本标注置信度，将最可信的无标签样本加入到原始标签样本集合，以扩大可用标签样本集的数量。本书以该算法结合Gkclust模糊聚类算法，将有标签和无标签的样本有效地结合起来，期望训练出一个效果更好的分类器，如图1-1(c)部分所示。

6. 分类器集成学习算法

集成学习首先训练多个有差异的学习器，然后对多个分类器的分类结果进行组合来决定最终的分类。相对于单个学习器，集成学习算法可以显著地提高学习系统的泛化能力。本书在研究半监督SVM分类器的同时，产生了若干个性能差异的半监督分类器个体，通过加权投票策略将这些半监督分类器个体有效地结合起来，期望可以进一步改善单个分类器的不足，更快地达到理想的分类精度，如图1-1(d)部分所示。

方法流程图如图1-1所示。

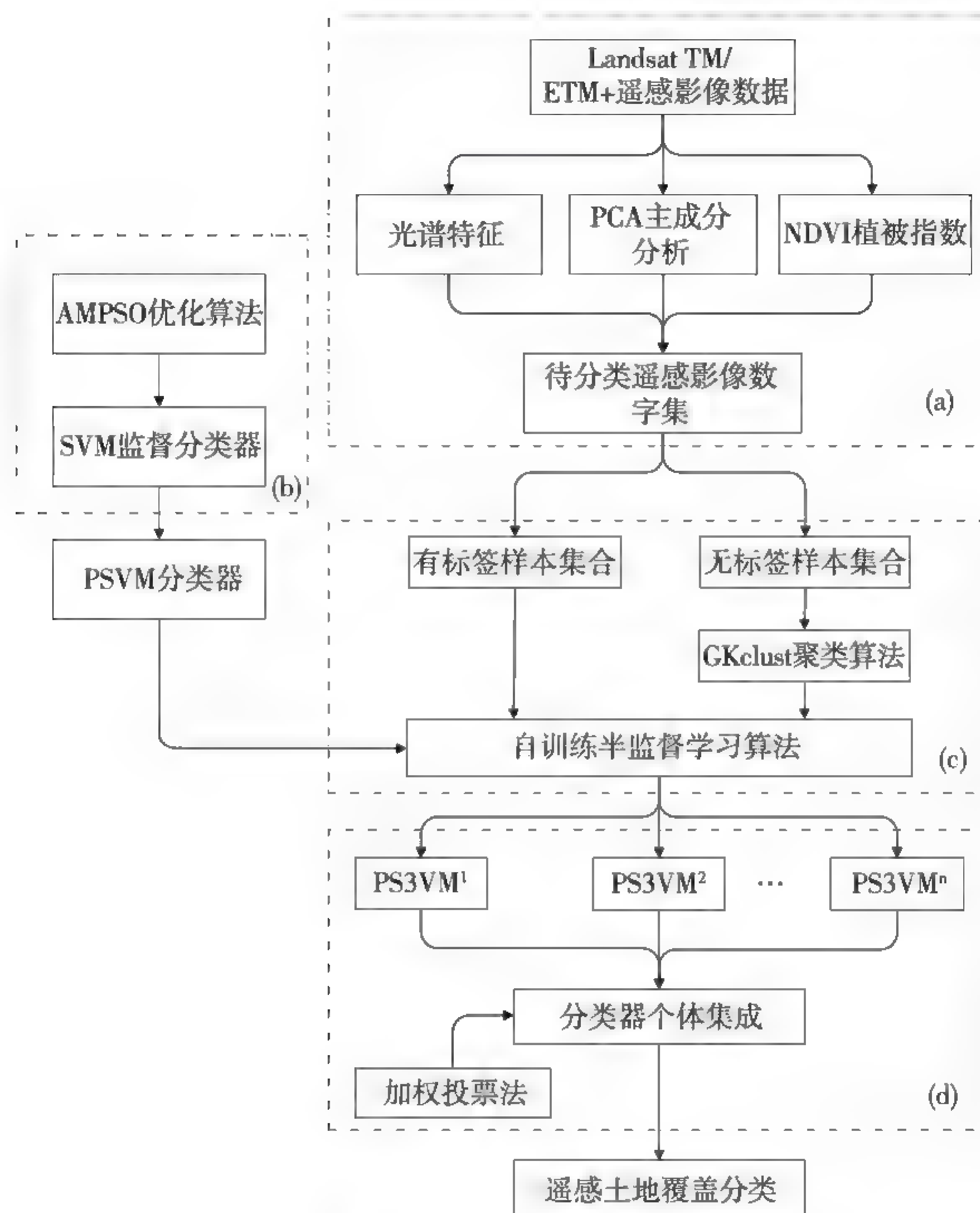


图1-1 方法流程图

1.3.2 结构安排

第1章 绪论

本章介绍机器学习、遥感分类、土地覆盖相关概念；目前机器分类在遥感应用领域存在的问题及研究意义；本研究的研究方法及主要

内容及创新点。

第2章 关键技术国内外研究现状

本章介绍SVM遥感分类研究进展；半监督学习、集成学习研究现状；目前遥感影像信息提取技术存在问题。

第3章 遥感图像数字化

本章确定研究区域，以行列号115-30一景、近20年的6幅不同时相的Landsat ETM/TM影像作为研究材料。讨论本书所采用的影像合成方式、特征采集方法、土地覆盖分类依据，以及特征选取方法，为进一步研究分类方法提供必要的数据材料。

第4章 SVM参数优化方法研究

本章针对传统PSO优化SVM参数存在早熟收敛、后期迭代效率不高从而造成参数寻优不准确缺点，提出了一种基于自适应变异粒子群算法的SVM参数优化模型(AMPSO-SVM)。并将所提方法与SVM和传统PSO优化SVM方法进行比较。

第5章 基于模糊聚类的半监督支持向量机土地覆盖分类方法研究

本章一方面利用自训练半监督算法，在较少训练标签样本的基础上，同时利用大量、廉价的未标记样本，目的在于挖掘未标记样本中所蕴涵固有结构信息，来对分类器偏差情况进行娇正；此外，为了避免自训练算法的“错误累积”现象，将以SVM作为基分类器，并将所提出的改进智能优化算法为SVM参数选择的依据。另一方面，由于遥感图像存在同物异谱、同谱异物、混合像元等特点，将模糊聚类算法引入到未标签样本标记过程中。

第6章 基于半监督集成支持向量机的土地覆盖分类研究

本章从个体生成(使用程序来生成个体分类器)和结论生成(选择

特定的策略来组合分类器)两个部分考虑，提出半监督集成SVM分类策略。该策略分两个步骤：首先，个体生成部分一方面利用改进智能算法优化SVM分类器参数以获得高精度分类器个体，另一方面采用Gustafson-Kessel聚类算法控制自训练算法错误标记样本的加入以提高个体分类器的差异性；然后，结论生成部分采用加权投票策略将半监督分类器个体集成，最后将该模型应用于土地覆盖遥感分类实验。

第7章　总结与展望

本章对专著主要研究内容进行总结，深入分析遥感分类领域的研究方法现状及不足，探讨新的土地覆盖遥感分类方法。同时，陈述本书研究的不足及对未来研究进行展望。

参考文献

[1] 唐根年. 区域土地利用/土地覆被变化动态监测与生态影响评价研究[D]. 浙江大学, 2002.

[2] 赵伟. 基于遥感的土地利用覆盖变化及其生态环境效应研究——以重庆市“一小时经济圈”[D]. 西南大学, 2008.

[3] 朱蕾. 土地利用/覆盖变化及其对生态安全的影响研究[D]. 浙江大学, 2007.

[4] 那晓东, 张树清, 孔博. 三江平原土地利用覆被动态变化对洪河保护区湿地植被退化的影响[J]. 干旱区资源与环境, 2009, 23(3): 144-150.

[5] 刘殿伟. 过去50 年三江平原土地利用/覆被变化的时空特征与环境效应[D]. 吉林大学, 2006.

[6] Giertz, S., Junge, B., Diekkruger, B.. *Assessing the Effects of Land Use Change on Soil Physical Properties and Hydrological Processes in the Sub-humid Tropical Environment of West Africa*[J]. Physics and Chemistry of the Earth, 2005, 30: 485-496.

[7] 郭旭东. 河北省遵化县土地利用与土壤养分变化[D]. 中国科学院生态环境研究中心, 2001.

[8] Zhang Q.Y., Li F.D., Liu M.Y.. *Effect of Land Use on Soil Properties in Debris Flow Bottomland: A Case Study at Xiaojiang Basin, Yunnan*[J]. Wuhan University Journal of Natural Sciences, 2006, 11(4): 870-874.

[9] Zhao M., Zeng X.M.. *A Theoretical Analysis on the Local Climate Change Induced by the Change of Land Use*[J]. Advances in Atmospheric Sciences, 2002, 19(1): 45-63.

[10] Wang H.J., Shi W.L., Chen X.H.. *Change Caused by Land Use and Land Cover Variation in West China*[J]. Advances in Atmospheric Sciences, 2006, 23(3): 355-364.

[11] Chen Y.H., Shi P.J., Li X.B.. *A Combined Approach for Estimating Vegetation Cover in Urban/Suburban Environments from Remotely Sensed Data*[J]. Computers & Geosciences, 2006, 32: 1299-1309.

[12] Kozlowski, T.T.. *Responses of Woody Plants To Human-Induced Enviromental Stresses: Issues, Problems, and Strategies for*

Alleviating Stress[J]. Critical Reviews in Plant Science, 2000, 19(2): 91-170.

[13] Xu C., Liu M., An S.. *Assessing the Impact of Urbanization on Regional Net Primary Productivity in Jiangyin County, China*[J]. Journal of Environmental Management, 2007, 85(3): 597-606.

[14] Feng X., Liu G., Chen J.M., Chen M., Liu J., Ju W.M., Sun R., Zhou W.. *Net Primary Productivity of China's Terrestrial Ecosystems from a Process Model Driven by Remote Sensing*[J]. Journal of Environmental Management, 2007, 85(3): 563-573.

[15] 马晓薇. 黄河流域土地利用变化及其对径流泥沙的影响[D]. 北京师范大学, 2003.

[16] Ranjan S.P., Kazama S., Sawamoto M.. *Effects of Climate and Land Use Changes on Groundwater Resources in Coastal Aquifers*[J]. Journal of Environmental Management, 2006, 80: 25-35.

[17] Samaniego L., Bardossy A.. *Simulation of the Impacts of Land Use/Cover and Climatic Changes on the Runoff Characteristics at the Mesoscale*[J]. Ecological Modeling, 2006, 196: 45-61.

[18] 袁艺, 谢锋, 史培军. 快速城市化过程中城镇用地与农业用地的景观斑块特征研究——以深圳市为例[J]. 北京师范大学学报(自然科学版), 2003(6).

[19] Roo A.D., Schmuck G., Perdigao V.. *The Influence of Historic Land Use Changes and Future Planned Land Use Scenarios on Floods in the Oder Cathment*[J]. Physics and Chemistry of the Earth, 2003, 28: 1291-1300.

[20] 叶树华, 任志远. 遥感概论[M]. 陕西: 陕西科学技术出版社, 1993.

[21] 彭望碌. 遥感概论[M]. 北京: 高等教育出版社, 2002.

[22] 孙家柄, 舒宁, 关泽群. 遥感原理、方法和应用[M]. 北京: 测绘出版社, 2002.

[23] Serge A., Ludovic R., Yannick C., Alain B.. *A Fuzzy-possibilistic Scheme of Study for Objects with Indeterminate Boundaries:Application to French Polynesian Reefscapes*[J]. IEEE Transaction on Geoscience and Remote Sensing, 2000, 38(1): 257-270.

[24] Yang C., Bruzzone L., Sun F.Y., Lu L.J., Guan R.C., Liang Y.C.. *A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in Multispectral Images*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48 (6): 2647-2659.

[25] 张睿, 马建文. 支持向量机在遥感数据分类中的应用新进展[J]. 地球科学进展, 2009, 24(5): 555-562.

[26] 杨晨. 基于机器学习的土地覆盖遥感信息提取方法研究[D]. 吉林大学, 2010.

[27] 张东晖，等译. 生物信息学——机器学习方法[M]. 北京：中信出版社, 2003.

[28] Leung Y., Fung T., Mi J.S.. *A Rough Set Approach to the Discovery of Classification Rules in Spatial Data*[J]. International Journal of Geographical Information Science, 2007, 21 (9): 1033-1058.

[29] Tissari S., Nykänen V., Lerssi J., Kolehmainen M.. *Classification of Soil Groups Using Weights-of-evidence Method and*

RBFLN-neural Nets[J]. Natural Resources Research, 2007, 16 (2): 159-169.

[30] Vapnik V.N.. *Statistical Learning Theory*[M]. New York: Wiley, 1998.

[31] Foody G.M., Mathur A.. *A Relative Evaluation of Multiclass Image Classification by Support Vector Machines*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42 (6):1335-1343.

[32] Zhang R., Ma J.. *Feature Selection for Hyperspectral Data based on Recursive Support Vector Machines*[J]. International Journal of Remote Sensing, 2009, 30 (14): 3669-3677.

[33] Wang L., Jia X.. *Integration of Soft and Hard Classifications using Extended Support Vector Machines*[J]. IEEE Geoscience and Remote Sensing Letters, 2009, 6 (3): 543-547.

[34] Tuia D., Pacifici F., Kanevski M., Emery W.J.. *Classification of Very High Spatial Resolution Imagery Using Mathematical Morphology and Support Vector Machines*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47 (11): 3866-3879.

[35] Petropoulos G.P., Kalaitzidis C., Vadrevu K.P.. *Support Vector Machines and Object-based Classification for Obtaining Land-use/cover Cartography from Hyperion Hypspectral Imagery*[J]. Computers & Geosciences, 2012, 41: 99-107.

[36] Liu Y., Zhang B., Huang L.H., Wang L.M.. *A Novel Optimization Parameters of Support Vector Machines Model for the Land use/cover Classification*[J]. International Journal of Food, Agriculture & Environment, 2012, 10 (2): 1098-1104.

[37] Knorn J., Rabe A., Radeloff V.C., Kuemmerle T., Kozak J., Hostert P.. *Land Cover Mapping of Large Areas Using Chain Classification of Neighboring Landsat Satellite Images*[J]. Remote Sensing of Environment, 2009, 113(5): 957-964.

[38] Heikkinen V., Tokola T., Parkkinen J., Korpela, I., Jaaskelainen T.. *Simulated Multispectral Imagery for Tree Species Classification Using Support Vector Machines*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(3): 1355-1364.

[39] Lardeux C., Frison P.L., Tison C., Souyris J.C., Stoll B., Fruneau B., Rudant J.P.. *Support Vector Machine for Multifrequency SAR Polarimetric Data Classification*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(12): 4143-4152.

[40] Huang X., Zhang L.P.. *Road Centreline Extraction from High-resolution Imagery based on Multiscale Structural Features and Support Vector Machines*[J]. International Journal of Remote Sensing, 2009, 30 (8): 1977-1987.

[41] 冯晓毅, 王西博, 王雷. 基于改进JSEG算法的高分辨率遥感图像分割方法[J]. 计算机科学, 2012, 39(8): 284-287.

第2章

关键技术国内外研究现状

2.1 遥感影像信息提取方法

遥感分类是获取遥感影像信息的一个重要环节，早在20世纪50年代，人们就开始探讨利用遥感资料进行大范围制图的可行性，并发展了适用于遥感数据特点的相关研究地物分类系统及其分类方法[1]。其中，传统分类方法是基于地形图、调查数据以及一些航空图片的方法对地物变化进行地面测量。然而这种方法不仅速度慢，而且周期长，不能达到时时、动态监测的需要。近几十年来，随着传感器技术的发展，遥感影像波谱分辨率、空间分辨率不断提高，同时计算机技术、地理信息系统等相关理论技术不断进步，许多学者对遥感影像的自动分类算法进行了大量有益的研究，包括监督分类学习方法，如基于统计分析分类方法(最大似然法，朴素贝叶斯分类)、决策树分类方法、神经网络分类方法，非监督分类学习方法，如K-means算法、模糊C-means算法、AP聚类算法等。

基于传统统计分析法的土地覆盖分类是目前较为成熟、应用较多的算法，如最大似然分类法(Maximum Likelihood Classification，MLC)。MLC假设同一类别的训练集中各点的整体分布属于正态分布，利用训练集可求出方差、协方差以及均值等特征参数，从而求出总体的先验概率密度函数。骆剑承等[2]提出基于有限混合密度理论的期望最大(EM)算法来作为最大似然函数(MLC)参数估计的方法构建EM2MLC模型，应用于土地覆盖分类，并将分类结果进行了定性和定量的综合比较，认为EM2MLC在精度上得到了提高。李全等[3]分析土地覆盖分类应用中最大似然分类算法的不足，包括确定分类类别及数量、光谱波段、先验概率、选择和改进样本区，

以南京市TM数据为例进行了土地覆盖分类实验，结果证明此方法能有效提高分类精度。最大似然分类法的扩展就是贝叶斯分类法(Bayesian Classification)，这种方法在概率估计时增加了加权因子进行判断，由于它具有综合先验信息和数据样本信息的能力，而且简单有效，所以得到了广泛的应用[4]。最大似然分类法的主要缺点是，在对每一个像元分类时需要进行大量的计算来判断众多的光谱通道和鉴别多种光谱类型；此外，基于统计学分类器以变量之间“独立性假设”为前提，而在现实世界中，这种假设往往很难被满足。

决策树方法是多元统计分类的一种方法。决策树算法用于遥感分类的优势在于对数字影像数据特征空间的分割，主要表现为分类结构简单明了，对数据特征空间分布不需要预先假设某种参数化密度分布。并且它属于严格“非参”分类方法，对于输入数据空间特征和分类标识具有更好的弹性和稳健性。

Pal和Mather[5]提出利用决策树方法对不同的研究区域以及不同传感器所获得的影像进行土地覆盖分类，充分研究训练集尺寸、特征维数，以及特征选择对分类器分类精度的影响。

潘琛等[6]结合遥感图像分类的特点，探讨了决策树分类的实现方法和关键问题，以徐州市TM影像作为数据源进行了分类试验。试验说明了决策树分类法在遥感图像处理中的具体实现过程，并且试验结果表明该方法在依据感兴趣区类别进行图像分类时效果较好。

那晓东等[7]利用TM卫星影像数据，基于半方差分析和Z检验方法遴选最优的窗口大小、纹理特征及其派生波段，采用快速、无偏、高效统计树算法集成遥感影像的光谱特征、多尺度纹理特征和地学辅助数据，建立研究区湿地信息提取的决策树模型。结果表明

分类精度较好，是内陆淡水沼泽湿地信息提取的有效手段。决策树算法的不足在于它需要大量的训练样本来探究各类别属性间的关系，而且算法基础比较复杂，当空间数据特征比较简单且样本量不足时，分类结果表现较差。

神经网络属于非参数分类器，该方法被用于遥感分类，始于1988年。与传统统计分析方法相比，神经网络分类方法不需要预定义分类中各个数据源的先验权值，也不需要任何关于统计分布的先验知识。因此，它用于遥感影像分类时不必考虑像元统计分布特征。

Kavzoglu和 Mather[8]提出使用两个数据集优化的神经网络参数，将其分类模型用于土地覆盖分类实验，取得很好的分类效果。

熊桢等[9]将人工神经网络技术和分层处理技术结合，设计了一种分层神经网络分类算法，并将其用于土地覆盖分类。实验表明这种分层神经网络分类算法可以明显地提高分类精度，并对不规则分布的复杂数据具有很强的处理能力。

王任华等[10]应用人工神经网络模型对陆地卫星TM多光谱图像进行了森林植被分类的研究，所采用的网络模型为3层误差后向传播神经网络模型，实验还引入高程数据作为一个独立波段与3个多光谱波段一起进行分类，效果明显，对存在同物异谱现象的阔叶林分类精度也有一定程度的提高。尽管如此，神经网络也存在着局部最小化、结构参数选取困难以及过学习等问题，在具体应用过程中不得不耗费大量时间进行参数选择、样本预处理等工作。

上述分类方法均基于监督分类，又称为训练区分类，它对充足且准确的训练样本数量有很强的依赖性。非监督聚类，作为遥感分类的另一种方法，不需要预先定义训练样本，仅凭遥感图像地物

的光谱特征分布规律自然地进行分类，最后通过对各类光谱响应曲线的进一步分析得出类别属性。其中，K-means算法因其收敛速度快、算法结构简单已成为遥感分类的常用聚类算法之一[11]，但是K-means属于硬聚类算法，每个待分像元必须严格划分到各个类别中，而遥感影像的随机性及不确定性使得混合像元问题难以进行非此即彼的划分；此外，K-means算法容易陷入局部极小值，对异常值敏感。针对这些问题，许多学者提出利用模糊集理论为地物进行软划分来解决遥感数据混合像元的分类，尤淑撑等[12]将模糊分类技术用于多时相的Scan SAR的作物识别，认为其与传统的最大似然分类法相比，具有较高的识别精度；Wang和Mo[13]将模糊集理论应用于遥感图像分类；Thitimajshima[14]将Fuzzy c-means聚类算法应用于多光谱图像分割取得很好的结果。

除了K-means和Fuzzy c-means算法，Yang[15]提出将快速高效吸引子传播算法(Affinity Propagation，AP)与模糊统计学相似性度量的概念结合，提出了基于模糊统计学的吸引子传播算法(Fuzzy Statistics-Based Affinity Propagation，FS-AP)，并将其应用到多光谱遥感图像分类中，提高了分类的精度和效率。

以上研究表明，在遥感影像分类领域，各类算法各具特点，新的分类技术相比于传统方法在分类精度上有了一些提高，但也存在一定的不足。因此，继续探索先进方法以提高遥感影像的分类精度仍然存在广阔的研究空间。

2.2 SVM遥感分类研究进展

SVM是一种基于统计学习理论的新型监督学习算法，通过对偶优化形式在高维特征空间中寻找最优分类超平面，从而解决复杂数据的分类及回归问题。SVM算法具有结构简单，泛化能力强，易解决具有高维特征、小样本与不确定性等问题的优势[16]；能够有效克服统计方法所要求特征向量服从正态分布的要求，同时也解决了神经网络方法中无法避免的局部极值问题，非常适合高维、复杂的小样本多维数据分类。

2.2.1 SVM在遥感分类中的优点

1. SVM是一种稳定的分类器，具有较好的泛化能力

SVM理论适合有限样本分类，即其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优解。

Foody和Mathur[17]利用SVM进行遥感影像分类，仅仅使用SPOT HRV卫星影像的四分之一作为原始训练样本就可以得到很高的分类精度。

Sahoo等[18]使用SVM算法进行地质分类的实验，实验结果表明，SVM适用于小样本数据分类，并且表现出很强的鲁棒性。

Mathur and Foody[19]提出Crop Land Mapping Equivalent方法能有效降低样本数量。

Muñoz-Marí等[20]用单类分类器对单类和多类遥感问题分类并进行比较，结论指出，作为支持向量域描述(SVDD)的单类分类器，在处理不完整数据集时，表现得尤为出色。

2. SVM在处理高维的高光谱遥感数据表现性能更优

高光谱遥感图像能够提供几乎连续的地物波谱曲线，使得高光谱数据具有维数高、数据量大、数据不确定性等特点。SVM算法通过核函数将非线性变换映射到高维的特征空间(Feaurte Space)，在高维空间中构造线性判别函数来实现原空间中的非线性判别函数，巧妙地解决了维数问题。

Melgani和Bruzzone[21]将SVM应用于高光谱遥感图像分类，并与K-近邻分类器和RBF神经网络进行对比，从精度、稳定性、计算速度及计算复杂性方面进行论证，说明了SVM应用于高光谱图像分类的有效性。

谭琨和杜培军[22]从SVM基本理论出发建立了一个基于SVM的高光谱分类器，分析比较了各种SVM核函数对高光谱遥感图像分类精度的影响，利用网格搜寻的方法来确定SVM参数值，结果表明SVM进行高光谱分类的时候径向基核函数是分类的首选。

Chen等[23]提出了一个由多层次堆叠的SVM改进分类模型，该方法还包含有效区分两个特征空间的信息(如大小和形状)。实验表明，模型具有很强的泛化能力，同时多特征空间的利用对高光谱图像分类精度的提高也是有效的。

3. SVM可以获得全局最优解，避免局部极值问题

SVM算法将训练问题转化为一个求解受约束的二次型规划问题，从理论上讲得到的将是全局最优点，由此可以避免神经网络训练结果不稳定、容易陷入局部极小的问题[24]。此外，SVM易处理高维数据的优势，也使得多源遥感数据所组成的高维特征空间分类不会出现传统统计分类方法所不可避免的过学习问题[25]。

Huang等[26]将SVM应用于多光谱遥感土地利用分类，并与最大

似然法、人工神经网络和决策树方法进行对比，论证了SVM在精度与稳定性方面都优于以上方法。

骆剑承等[27]以SPOT全色波段影像上城市特征信息的提取为应用实例，并与人工神经网络等特征提取方法进行综合比较，认为SVM方法不但能够获得比较高的分类精度，而且在自适应能力、学习速度、特征空间高维不限制、可表达性等方面具有优势。

2.2.2 SVM在遥感影像分类中的不足

虽然SVM在遥感图像分类中取得了很好的效果，但仍存在有待改进和完善之处，至少在以下几个问题方面还需要进行深入研究。

1. 核函数和分类参数(包括惩罚系数c、核函数参数)的选择没有特别好的办法，应用时不容易找到最优的核函数和分类参数

核函数的选择在SVM分类处理中发挥着重要作用，MERCER准则是构造核函数应该满足的条件，通常包括全局核和局部核，前者如线性核、多项式核和Sigmoid核函数等，后者如径向基核、KMOD核函数等。如何针对特定问题选择核函数，目前并无准确原则，是一个比较困难的问题。此外，设置合适的SVM核函数参数对提高遥感影像分类精度也是非常重要的。SVM的参数主要指的是惩罚参数c和核函数参数。惩罚系数c是一个重要的量，控制对错分样本的惩罚程度。核函数的参数通常影响决策面的复杂性。如果不适当地选取这两个参数，可能产生“过学习”或“欠学习”的问题[41]。

2. 遥感数据的模糊性、不确定性是影响SVM分类精度的原因之一

遥感数据的模糊性和不确定性表现在，遥感图像自身的像素值

是不准确的；遥感映射机制可能造成的歧义性；遥感图像地物的边界是模糊的[42]。因此，遥感图像存在同物异谱、同谱异物、混合像元等问题，在分类过程中对混合像元很难进行非此即彼的划分，这些很容易造成SVM分类器的错分、漏分现象，最终导致分类器精度低下。

3. SVM分类器依赖于充足且代表性好的样本

准确先验知识或充足的训练样本是保证监督分类器分类精度的重要条件。它们的获取一方面消耗大量的人力和财力，另一方面人为选择样本时对待分类影像认知的有限性以及选择时的盲目性等因素，均会导致得到的样本数量少且代表性不好[43]~[44]。当训练样本集数量远远小于测试样本集，即便SVM善于处理小样本问题，也难以保证取得理想的分类效果。

2.2.3　SVM在遥感影像分类中的应用领域

近年来，国内外许多学者将SVM应用于遥感影像的处理与分类中[28]~[30]。具体应用领域包括以下几个方面。

1. 土地覆盖分类

Pal[31]提出非穷尽搜索与遗传算法(Genetic Algorithms，GA)结合SVM的分类方法，并将其应用于土地覆盖分类任务。

Carrão 等[32]利用SVM分类方法对多光谱MODIS遥感数据进行处理，解决土地利用分类问题。

张锦水等[33]基于变化向量分析方法，将光谱与纹理两种信息复合计算变化强度，并采用SVM法提取变化/非变化信息，通过监督分类确定变化区域内的土地利用/覆盖类型的转移方向，完成土地利

用/覆盖动态监测。

2. 作物生长监测

Tan 等[34]利用熵分解和SVM相融合的分类技术对多时相SAR图像分类，以实现水稻监测。

顾幸生等[35]在标准最小二乘SVM的基础上，利用改进的粒子群算法优化SVM模型参数，提出了基于IPSO-LS-SVM的软测量建模方法，建立了作物叶水势软测量模型。

3. 森林植被分类

张友静等[36]利用高分辨率卫星影像IKONOS，以实验区与验证区城市植被类型信息为对象，基于常用的参数和非参数分类方法的对比，分析SVM的核函数及其参数对分类精度的影响程度，构建了SVM决策树的城市植被类型分类模型。

Dalponte等[37]利用SVM进行森林类型识别，取得很好的效果。

4. 城市增长变化监测

Nemmour and Chibani[38]研究利用Landsat影像进行城市变化监测，试验表明SVM方法优于神经网络方法。

Licciardi等[39]利用5种算法对高光谱影像数据分类以实现城市区域变化监测，结果表明SVM分类方法性能最优。

沈体雁等[40]以我国京津唐都市圈为试验区，采用MODIS遥感影像作为主要数据源，基于SVM分类技术，探讨实现一种新的城市空间增长和城市土地利用变化遥感信息提取方法。

2.3　半监督学习理论及研究进展

监督学习指学习机通过对大量有标签的训练样本进行学习，从而推导出一个预测模型，判断未知样本点的标签。无监督学习是在完全没有标签的样本集中学习，推导出样本数据集的结构。半监督学习方法介于二者之间，在利用有标签样本学习的同时，挖掘未标记样本中所蕴涵的大量可用信息，利用一些分布上的假设或者样本之间的内在联系，将未标签样本转化为有标签样本，然后合并到有标签的数据中，扩大可用的训练数据集，从而使分类器的性能更优异。在遥感信息提取中，通过少量的标记样本学习来完成大量未标记样本的自动分类，对于遥感图像处理具有重要的研究意义。

半监督学习是近年来随着统计学习技术的不断发展，以及利用无标签样本这一需求越来越强烈而广泛受到关注[45]。半监督学习思想产生于20世纪60年代，一般认为其真正的研究工作始于Shahshahani 和Landgrebe[46]所提出的半监督学习理论。充足且准确的训练样本是提高分类精度的重要条件之一，遥感影像本身具有复杂性、随机性等特点；且人为选取样本总是存在经验、知识有限和盲目选择等缺点，这些造成所选择的分类样本不足且代表性不好，如果仅使用少量“昂贵的”有标签样本而不利用大量“廉价的”未标签样本，则是对数据资源的极大浪费。因此，在标签样本较少时，如何利用大量的未标签样本来改善学习性能具有重要的理论价值和应用意义。不同的论著、不同的研究者对半监督涉及方法的研究方向有不同的划分。通常，按其工作方式可以大致分为以下几类。

1. 基于EM(Expectation-Maximization)的半监督学习算法

EM算法[47]~[48]是一种模型构造法，基于聚类假设，是半监督学习中提出的比较早的一种学习方法。EM算法利用无标签样本反复估计和修正有标签样本获得假设中的模型参数，直到模型中的参数不再变化或者几乎不再变化为止。

EM方法在遥感分类领域有以下局限性：

(1) 如果假设的模型与实际不符，将会导致较大的错误率。一般而言，在未标记样本数量足够大的时候可以得到符合影像中各类别地物像素实际分布的、较为合理的先验概率；但当其数量不是很大时，因未标记样本选择的随机性，影像也可能会出现分布较少类别未能选取到样本或样本数量极少而使其比例失衡，结果导致影像中分布较少的弱势类别无法被正确划分。

(2) 基于生成模型的分类中，混合成分之间的重叠也是影响最终分类精度的重要因素之一。

2. 基于协作训练(Co-training)半监督学习算法

协作学习法[49]~[51]隐含地利用了聚类假设或者流形假设，其主要思想是把样本划分为两个子集，且满足下述条件：一是每个特征子集都是充足的训练样本，即在每个特征子集上都足以训练一个好的分类器；二是对于指定类别标签，每个特征子集条件都独立于另一个特征子集，即这两个特征子集必须在给定类别的条件下独立分布。在学习过程中，首先根据两个子集分别训练两个独立的分类器，这两个学习器挑选若干个标记置信度高的未标记样本进行相互标记，从而使对方利用这些新标记的样本进行更新，这个过程不断迭代进行，直到满足某个停止条件。缺点是该方法对样本的特征空

间作了很强的假设。

3. 基于TSVM(Transductive Support Vector Machines)半监督学习算法

20世纪70年代中期出现了直推式学习方法的主要思想，后来由Joachims[52]将该思想结合SVM提出直推式支持向量机(Transductive SVM，TSVM)方法。TSVM是基于标准SVM的一个扩展来解决半监督学习问题的，它不是直接将无标签样本点加入到标准SVM的优化函数中得到最优解，而是试图不改变SVM的优化机理，逐步通过标准SVM的训练算法修改划分超平面，并交换超平面两侧某些未标签样本的可能标签，将无标签样本转化为有标签样本，使得SVM在所有训练数据(包括有标签和无标签样本点)上最大化间隔(Margin)，从而得到一个既通过数据相对稀疏的区域又尽可能正确划分有标签样本的超平面。目前，TSVM在遥感影像分类中还没有引起足够的重视，为数不多的研究成果中较有代表性的是Bruzzone等[53]~[54]以及Tuia和Camps-valls[55]的工作。该算法的一个弱点是需要事先确定属于某个类别未标记样本的具体数目，对于遥感影像中的地物分类，事先估计某个类别样本的具体比例或者数目，是不可行也不可能的，一旦错误估计，对分类结果的影响将是灾难性的。

4. 基于自训练(Self-training)半监督学习算法

自训练法[56]~[57]是半监督学习的一种比较常用的方法。自训练法首先利用有标签样本训练一个分类器，然后用此分类器对所有无标签样本进行分类，并给每个无标签样本标上相应的类别标签和置信度；再将置信度高的样本连同它的类别标签合并到训练集中继续训练分类器；重复上述过程直至结束条件满足。这种方法的缺点是在学习过程中，一旦某个分类出错，将导致这个错误被继续学习和加

强，即出现所谓的错误累积现象。为了避免这种情况的发生，通常为算法设置一个阈值，当未标注样本分类预期值小于该阈值时，算法便认为该样本为无效样本值，不将其加入到标注样本训练集中，以控制错误信息的输入。

2.4 半监督分类中的聚类算法

半监督分类中的聚类[58]~[61]作为一种从无标记样例中提取信息的方法，目的在于改进分类任务。遥感影像数据可被分类的前提是属于同一个类别的影像单元在特征空间中具有聚类的特点，且各类别之间在特征空间中的样本点是稀疏的，这种聚类的特点表现为一种全局的特征，同时满足基于生成模型和基于直推式学习的前提假设。

除了上述这些半监督学习方法外，还包括基于流形或图谱的一些图正则化框架的半监督学习方法[62]~[65]。

2.5 集成学习理论及研究进展

集成学习(Ensemble learning)也是当今机器学习的研究热点之一，弱强学习算法之间的转化思想是集成学习的理论基础，即Kearns和Valiant[66]所提出的是否可以将弱可学习算法提升为强可学习算法，如果两者等价，那么在学习概念时，只需要找到一个比随机猜测略好的弱可学习算法，通过某种形式的转换，就可以将其提

类、SVM、决策树等，由此产生了狭义、广义集成学习的概念。

狭义定义：集成学习是综合多个同构的学习机来对同一个问题进行学习。同构指集成中的所有程序学习机属于同一种类型，集成在某输入示例下的输出由构成集成的个体学习器在该示例下的输出共同决定。

广义定义：集成学习是用有限个学习器对同一个问题进行学习。学习器可以是任何类型，集成在某输入示例下的输出由构成集成的个体学习器在该示例下的输出共同决定。

集成学习直到20世纪90年代才逐步受到重视，在许多领域获得了优异的效果。如，Wolpert[71]提出Stacked Generalization算法；Perrone 和Cooper[72]提出了可以大幅提升回归估计性能的集成方法理路框架；Jordan和Jacobs[73]介绍了一个分级混合专家模型；Battiti和Colla[74]证明规模神经网络之间的错误相关性足够小，那么它们集成的性能也要大大超过最好的独立神经网络个体；Lam和Suen[75]比较了多种分类器组合方法，通过遗传算法获得权重，发现当训练集不完善时，简单的多数投票是最可靠的方式；Breiman[76]从可重复取样技术入手，提出了著名的Bagging算法；Freund和Schapire[77]改进了Boosting算法，进一步提出了自适应Boosting(AdaBoost)算法，目前AdaBoost已成为最流行的Boosting方法；Dietterich[78]发表了集成学习评论，将集成学习列为机器学习的四大研究方向的首位；Huang等[79]将神经网络集成用于图像在深度方向上发生偏转多姿态的人脸识别；Collobert等[80]通过将训练集划分为多个不相交的子集并在每个子集上构建个体SVM，最后将个体的预测结果加权平均作为集成的输出；Valentini等[81]采用特征选择方法解决维数灾难问

题，并采用Bagging技术克服由于小样本和生物差异性导致的方差问题；Liu等[82]构建随机C4.5决策树集成，获得了远好于支持向量机集成的分类结果；Hu等[83]和Mei等[84]利用粗糙集理论把输入特征空间分成多个特征子空间，运用各子空间训练SVM成员分类器，最后进行集成；Archibald and Fann[85]提出SVM的特征集成选择分类方法，取得准确的结果，并有效地降低了计算复杂度；Pal[86]提出多SVM分类器集成方法；Chen等[87]研究了成对决策树SVM的集成对高光谱数据分析，实验结果与One-against-oneSVM相似，这主要取决于基于决策树方法的层次结构；Ghoggali等[88]利用遗传算法和SVM集成利用有限可用训练样本对遥感数据分类，分类精度大幅提高；Mukhopadhyay和Maulik[89]将多目标模糊聚类与SVM集成进行非监督分类；邬俊等[90]提出一种基于偏袒性半监督集成的SVM主动反馈技术，在集成学习框架中使用未标记数据以增加个体分类器之间的差异性，还设计一种偏袒加权策略，使得集成分类模型对正样本给予更大的关注程度，以应对正负样本间的不对称分布问题；杨娜等[91]提出基于SVM无限集成学习方法，并将其应用于遥感图像分类，结果表明可显著提高遥感图像的分类精度。

以上集成方法思路分别侧重于两个步骤：首先使用特殊程序来生成个体分类器，然后选择特定的策略来组合分类器。

从以上论述不难看出，半监督学习和集成学习是机器学习两个重要方法，在过去几十年，二者取得了巨大成功，然而两个方法的发展几乎是并行的，只有少数研究涉及二者的结合。如何有效地将半监督学习、集成学习融合，并将其应用于SVM分类模型，是提高SVM分类器泛化能力的一个崭新思路。

参考文献

[1] 杨立明, 朱智良. 全球及区域尺度土地覆盖/土地利用遥感研究的现状和展望[J]. 自然资源学报, 1999, 14(4): 340-344.

[2] 骆剑承, 周成虎, 梁怡. 支撑向量机及其遥感影像空间特征提取和分类的应用研究[J]. 遥感学报, 2002, 6(1): 50-55.

[3] 李全, 王海燕, 李霖. 基于最大似然分类算法的土地覆盖分类精度控制研究[J]. 国土资源科技管理, 2005, 4: 42-45.

[4] Langley P., Sage S.. *Induction of Selective Bayesian Classifiers: From Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*[M]. Seattle, WA: Morgan Kaufmann, 1994.

[5] Pal M.. *Support Vector Machine-based Feature Selection for Land Cover Classification: a Case Study with DAIS Hyperspectral Data*[J]. International Journal of Remote Sensing, 2006, 27(14): 2877-2894.

[6] 潘琛, 杜培军, 张海荣. 决策树分类法及其在遥感图像处理中的应用[J]. 测绘科学, 2008, 22(1): 208-211.

[7] 那晓东, 张树清, 李晓峰, 于欢, 刘春悦. 基于Quest决策树兼容多源数据的淡水沼泽湿地信息提取[J]. 生态学杂志, 2009, 28(2): 357-365.

[8] Kavzoglu T., Mather P.M.. *The Use of Backpropagating Artificial Neural Networks in Land Cover Classification*[J]. International Journal of Remote Sensing, 2003, 24: 4907-4938.

[9] 熊桢, 郑兰芬, 童庆禧. 分层神经网络分类算法[J]. 测绘学报,

2000, 29(3): 229-234.

[10] E任华, 霍宏涛, 游先祥. 人工神经网络在遥感图像森林植被分类中的应用[J]. 北京林业大学学报, 2003, 25(4): 1-5.

[11] Ding Z. J., Yu J., Zhang Y.. *A New Improved K-means Algorithm with Penalized Term*[J]. IEEE International Conference Granular Computing, 2007: 313-317.

[12] 尤淑撑, 张玮, 严泰来. 模糊分类技术在作物类型识别中的应用[J]. 国土资源遥感, 2000, 1: 39-43.

[13] Wang Y., Mo J.. *Fuzzy Logic Applied in Remote Sensing Image Classification: In Proceeding International Conference Systems*[C]. Man and Cybernetics, 2004: 6378-6382.

[14] Thitimajshima P.. *A New Modified Fuzzy C-means Algorithm for Multispectral Satellite Images Segmentation*[J]. In Proceeding IGARSS, 2000, 4: 1684-1686.

[15] Yang C., Bruzzone L., Sun F.Y., Lu L.J., Guan R.C., Liang Y.C.. *A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in Multispectral Images*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(6): 2647-2659.

[16] Pal M., Mather P.M.. *Support Vector Machines for Classification in Remote Sensing*[J]. International Journal of Remote Sensing, 2005, 26(5): 1007-1011.

[17] Foody G.M., Mathur A.. *Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification*[J]. Remote Sensing of Environment, 2004, 93

(1-2): 107-117.

[18] Sahoo B.C., Oommen T., Misra D., Newby G.. *Using the One-dimensionals-transform as a Discrimination Tool in Classification of Hyperspectral Images*[J]. Canadian Journal of Remote Sensing, 2007, 33 (6): 551-560.

[19] Mathur A., Foody G.M.. Crop *Classification by Support Vector Machine with Intelligently Selected Training Data for an Operational Application*[J]. International Journal of Remote Sensing, 2008, 29(8): 2227-2240.

[20] Muñoz-Marí J., Bruzzone L., Camps-Valls G.. *A Support Vector Domain Description Approach to Supervised Classification of Remote Sensing Images*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2007, 45(8): 2683-2692.

[21] Melgani F., Bruzzone L.. *Classification of Hyperspectral Remote-sensing Images with Support Vector Machines*[J]. IEEE Transactions Geoscience and Remote Sensing, 2004, 42(8): 1778-1790.

[22] 谭琨, 杜培军. 基于支持向量机的高光谱遥感图像分类[J]. 红外与毫米波学报, 2008, 27(2): 421-423.

[23] Chen J., Wang C., Wang R.. *Using Stacked Generalization to Combine SVMs in Magnitude and Shape Feature Spaces for Classification of Hyperspectral Data*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(7): 2193-2205.

[24] 李晓宇, 张新峰, 沈兰荪. 支持向量机(SVM)的研究进展[J]. 测控技术, 2006, 25(5): 7-12.

[25] 张睿, 马建文. 支持向量机在遥感数据分类中的应用新进展[J]. 地球科学进展, 2009, 24(5): 555-562.

[26] Huang C., Davis L., Townshend J.. *An Assessment of Support Vector Machines for Land Cover Classification*[J]. International of Remote Sensing, 2002, 23(4): 725-749.

[27] 骆剑承, 王钦敏, 马江洪, 周成虎, 梁怡. 遥感图像最大似然分类方法的EM改进算法[J]. 测绘学报, 2002, 31(3): 234-239.

[28] Mountrakis G., Im J., Ogole C.. *Support Vector Machines in Remote Sensing: A Review*[J]. Photogrammetry and Remote Sensing, 2011, 66: 247-259.

[29] Zhang L., Huang X., Huang B., Li P.. *A Pixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(10): 2950-2961.

[30] Yu L., Porwal A., Holden E.J., Dentith M.C.. *Towards Automatic Lithological Classification from Remote Sensing Data Using Support Vector Machines*[J]. Computers & Geosciences, 2012, 45: 229-239.

[31] Pal M.. *Support Vector Machine-based Feature Selection for Land Cover Classification: A Case Study with DAIS Hyperspectral Data*[J]. International Journal of Remote Sensing, 2006, 27(14): 2877-2894.

[32] Carrão H., Gonçalves P., Caetano M.. *Contribution of Multispectral and Multitemporal Information from MODIS Images to*

Land Cover Classification[J]. Remote Sensing of Environment, 2008, 112(3): 986-997.

[33] 张锦水, 潘耀忠, 韩立建, 苏伟, 何春阳. 光谱与纹理信息复合的土地利用覆盖变化动态监测研究[J]. 遥感学报, 2007, 11(4): 500-510.

[34] Tan C.P., Koay J.Y., Lim K.S., Ewe H.T., Chuah H.T.. *Classification of Multitemporal Sar Images for Rice Crops Using Combined Entropy Decomposition and Support Vector Machine Technique*[J]. Progress in Electromagnetics Research, 2007, 71: 19-39.

[35] 顾幸生, 潘晔, 卢胜利. 基于改进支持向量机的作物叶水势软测量建模[J]. 同济大学学报:自然科学版, 2010, 11: 1669-1674.

[36] 张友静, 高云霄, 黄浩, 任立良. 基于SVM决策支持树的城市植被类型遥感分类研究[J]. 遥感学报, 2006, 2: 191-196.

[37] Dalponte M., Bruzzone L., Gianelle D.. *Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46 (5): 1416-1427.

[38] Nemmour H., Chibani Y.. *Multiple Support Vector Machines for Land Cover Change Detection: An Application for Mapping Urban Extensions*[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2006, 61 (2): 125-133.

[39] Licciardi G., Pacifici F., Tuia D., Prasad S., West T., Giacco F., Thiel C., Inglada J., Christophe E., Chanussot J., Gamba P.. *Decision Fusion for the Classification of Hyperspectral Data: Outcome of the*

2008 GRS-S Data Fusion Contest[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(11): 3857-3865.

[40] 沈体雁, 王煌基, 刘良明. 基于SVM和MODIS的城市增长遥感监测研究[D]. 第十五届全国遥感技术学术交流会论文摘要集, 2005.

[41] Jae H.M., Young C.L.. *Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters*[J]. Expert Systems with Applications, 2005, 28: 603-614.

[42] 承继成, 郭华东, 史文中. 遥感数据的不确定性问题[M]. 北京: 科学出版社, 2004.

[43] Chi M., Feng R., Bruzzone L.. *Classification of Hyperspectral Remote-sensing Data with Primal SVM for Small-sized Training Dataset Problem*[J]. Advances in Space Research, 2008, 41(11): 1793-1799.

[44] Foody G.M., Mathur A.. *The Use of Small Training Sets Containing Mixed Pixels for Accurate Hard Image Classification: Training on Mixed Spectral Responses for Classification by a SVM*[J]. Remote Sensing of Environment, 2006, 103(2): 179-189.

[45] 周志华. 半监督学习专刊前言[J]. 软件学报, 2008, 19(11): 2789-2790.

[46] Shahshahani B., Landgrebe D.. *The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon*[J]. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 1087-1095.

[47] Nigam K., Ghani R.. *Analyzing the Effectiveness and*

Applicability of Co-training[C]. Proceedings of the ninth international conference on Information and knowledge management, 2000: 86-93.

[48] Baluja S.. *Probabilistic Modeling for Face Orientation Discriminition: Learning from Labeled and Unlabeled Data*[C]. Neural Information Processing Systems, 1998: 854-860.

[49] Blum A., Mitchell T.. *Combining Labeled and Unlabeled Data with Co-training*[C]. In Proceedings of Computer Learning Theory, 1998: 92-100.

[50] Zhou Z.H., Li M.. *Semi-supervised Learning with Co-training*[C]. Edinburgh, Scotland: In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI' 05), 2005: 908-913.

[51] Zhou Z.H., Li M.. *Semisupervised Regression with Cotraining-style Algorithms*[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(11): 1479-1493.

[52] Joachims T.. *Making Large Scale SVM Learning Practical*[D]. Universität Dortmund, 1999.

[53] Bruzzone L., Chi M., Marconcini M.. *Transductive SVMs for Semisupervised Classification of Hyperspectral Data*[J]. Geoscience and Remote Sensing Symposium, IGARSS, 2005, 1: 164-167.

[54] Bruzzone L., Chi M., Marconcini M.. *A Novel Transductive SVM for Semisupervised Classification of Remote-sensing Images*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(11): 3363-3373.

[55] Tuia D., Camps-valls G.. *Semi-supervised Remote Sensing Image Classification with Cluster Kernels*[J]. IEEE Geoscience and Remote Sensing Letter, 2009, 6(2): 224-228.

[56] Rosenberg et al. *Semi-supervised Self-training of Object Detection Models*[J]. In Seventh IEEE workshop on applications of computer vision, 2005, 1: 29-36.

[57] Yarowsky D.. *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods*[C]. In Proceedings of the 33rd Annual meeting of the Association for computational Linguistics, 1995: 189-196.

[58] Fung G., Mangasarian O.L.. *Semi-Supervised Support Vector Machines for Unlabeled Data Classification*[J]. Optim. Methods Software, 2001, 15(1): 29-44.

[59] Li M., Cheng Y., Zhao H.. *Unlabeled Data Classification via Support Vector Machine and k-means Clustering*[C]. In Proceedings of the International Conference on Computer Graphics, Imaging and Visualization, CGIV' 04, 2004: 183-186.

[60] Zeng H. J., Wang X. H., Chen Z., Lu H. J.. *CBC: Clusteringbased Text Classification Requiring Minimal Labeled Data*[C]. Third IEEE International Conference on Data Mining, 2003: 443-450.

[61] Chapelle O., Weston J., Scholkopf B.. *Cluster Kernels for Semi-Supervised Learning*. In NIPS, 15.

[62] Zhou D.Y., Scholkopf B.. *Learning from Labeled and Unlabeled Data Using Random Walks*[C]. In Proceedings of 26th DAGM Symposium Pattern Recognition, Berlin: Springer-Verlag, 2004: 237-244.

[63] Belkin M., Niyogi P.. *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*[J]. Neural Computation, 2003, 15 (6): 1373-1396.

[64] Zhu X., Ghahramani Z., Lafferty J.. *Semi-supervised Learning Using Gaussian Fields and Harmonic Functions*[C]. In Proceedings of the 20th International Conference on Machine Learning (ICML-2003), Washington DC, 2003: 912-919.

[65] Blum A., Chawla S.. *Learning from Labeled and Unlabeled Data Using Graph Mincuts*[C]. In Proceedings of the 18th International Conference on Machine Learning (ICML' 01), San Francisco, CA, 2001: 19-26.

[66] Kearns M., Valiant L.G.. *Learning Boolean Formulae or Factoring*[D]. Technical Report TR-1488, Havard University, Cambridge, MA, 1988.

[67] 周志华, 陈世福. 神经网络集成[J]. 计算机学报, 2002, 25(1): 1-8.

[68] Dietterich T.G.. *Ensemble Methods in Machine Learning*[J]. In Multiple Classier Systems, Cagliari, Italy, 1857: 1-15.

[69] Hansen L.K., Salamon P.. *Neural Network Ensembles*[J]. IEEE Transactions on Pattern Analysis and Machine Intellignece, 1990, 12(10): 993-1001.

[70] Sollich P., Krogh A.. *Learning with Ensembles: How over-fitting Can Be Useful*[J]. In Advances in Neural Information Processing Systems, 1996, 8: 190-196.

[71] Wolpert D.H.. *Stacked Generalization*[J]. Neural Networks,

1992, 5: 241-259.

[72] Perrone M.P., Cooper L.N.. *When Networks Disagree: Ensemble Methods for Hybrid Neural Networks*[C]. Neural Networks for Speech and Image Processing, London: Chapman-Hall, 1993: 126-142.

[73] Jordan M.I., Jacobs R.A.. *Hierarchical Mixtures of Experts and the EM Algorithm*[C]. In Proceedings of 1993 International Joint Conference on Neural Networks, 1993, 2: 1339-1344.

[74] Battiti R., Colla A.M.. *Democracy in Neural Nets: Voting Schemes for Classification*[J]. Neural Networks, 1994, 7(4): 691-707.

[75] Lam L., Suen C.Y.. *Optimal Combination of Pattern Classifiers*[J]. Pattern Recognition Letters, 1995, 6(9): 945-954.

[76] Breiman L.. *Bagging Predictors*[J]. Machine Learning, 1996, 24(2): 123-140.

[77] Freund Y., Schapire R.E. *Experiments with a New Boosting Algorithm*[C]. In Machine Learning: Proceedings of 13th International Conference, 1996: 148-156.

[78] Dietterich T.G.. *Machine Learning Research: Four Current Directions*[J]. AI Magazine, 1997, 18(4): 97-136.

[79] Huang F.J., Zhou Z.H., Zhang H.,J., Chen T.. *Pose Invariant Face Recognition*[C]. In Proceedings of the IEEE International Conference on Neural Networks, 2000: 245-250.

[80] Collobert R., Bengio S., Bengio Y.. *A Parallel Mixture of Svms for Very Large Scale Problems*[J]. Neural Computation, 2002, 14(5): 1105-1114.

[81] Valentini G., Muselli M., Ruffino F.. *Bagged Ensembles of Support Vector Machines for Gene Expression Data Analysis*[C]. In Proceedings of the IEEE International Conference on Neural Networks, 2003: 1844-1849.

[82] Liu X.M., Hall O., Bowyer K.W.. *Comments on a Parallel Mixture of Svms for Very Large Scale Problems*[J]. Neural Computation, 2004, 16: 1345-1351.

[83] Hu Z.H., Cai Y.Z., Li Y., Xu X.M.. *Support Vector Machine Based Ensemble Classifier*[C]. Proceedings of the 2005 American Control Conference, 2005, 2: 745-749.

[84] Mei S.Y., Liu Y., Wu G.F., Zhang B.F.. *Rough Reducts Based Svm Ensemble*[C]. IEEE International Conference on Granular Computing, 2005: 571-574.

[85] Archibald R., Fann G.. *Feature Selection and Classification of Hyperspectral Images with Support Vector Machines*[J]. IEEE Geoscience and Remote Sensing Letters, 2007, 4(4): 674-677.

[86] Pal M.. *Ensemble of Support Vector Machines for Land Cover Classification*[J]. International Journal of Remote Sensing, 2008, 29(10): 3043-3049.

[87] Chen J., Wang C., Wang R.. *Combining Support Vector Machines with a Pairwise Decision Tree*[J]. IEEE Geoscience and Remote Sensing Letters, 2008, 5(3): 409-413.

[88] Ghoggali N., Melgani F., Bazi Y.. *A Multiobjective Genetic SVM Approach for Classification Problems with Limited Training*

Samples[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(6): 1707-1718.

[89] Mukhopadhyay A., Maulik U.. *Unsupervised Pixel Classification in Satellite Imagery Using Multiobjective Fuzzy Clustering Combined with SVM Classifier*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(4): 1132-1138.

[90] 邬俊, 段晶, 鲁明羽. 基于偏袒性半监督集成的SVM主动反馈方案[J]. 模式识别与人工智能, 2010, 23(6): 745-751.

[91] 杨娜, 秦志远, 张俊. 基于SVM无限集成学习方法的遥感图像分类[EB/OL]. http://www.cnki.net/kcms/detail/11.4415.P.20111230.1104.019.html, 2013.

第3章

遥感图像数字化

3.1 研究区位置及遥感影像集

3.1.1 研究区位置

本研究以图们江下游，中、朝、俄交界处作为研究对象。其中包括中国吉林、俄罗斯滨海边疆区Primorskiy kray，以及朝鲜咸镜北道Korea Hamgyeongbuk-do。其地理位置大约在41°06'～44°05'N 与127°39'～131°44'E 之间，如图3-1所示。该区地处温带大陆性季风气候区，冬季盛行西北风，夏季以东南风为主。气候的垂直变化明显，气温的年差较小，雨量比较充足，水、热同季。受日本海的影响，冬季比较暖和，夏季比较凉爽。

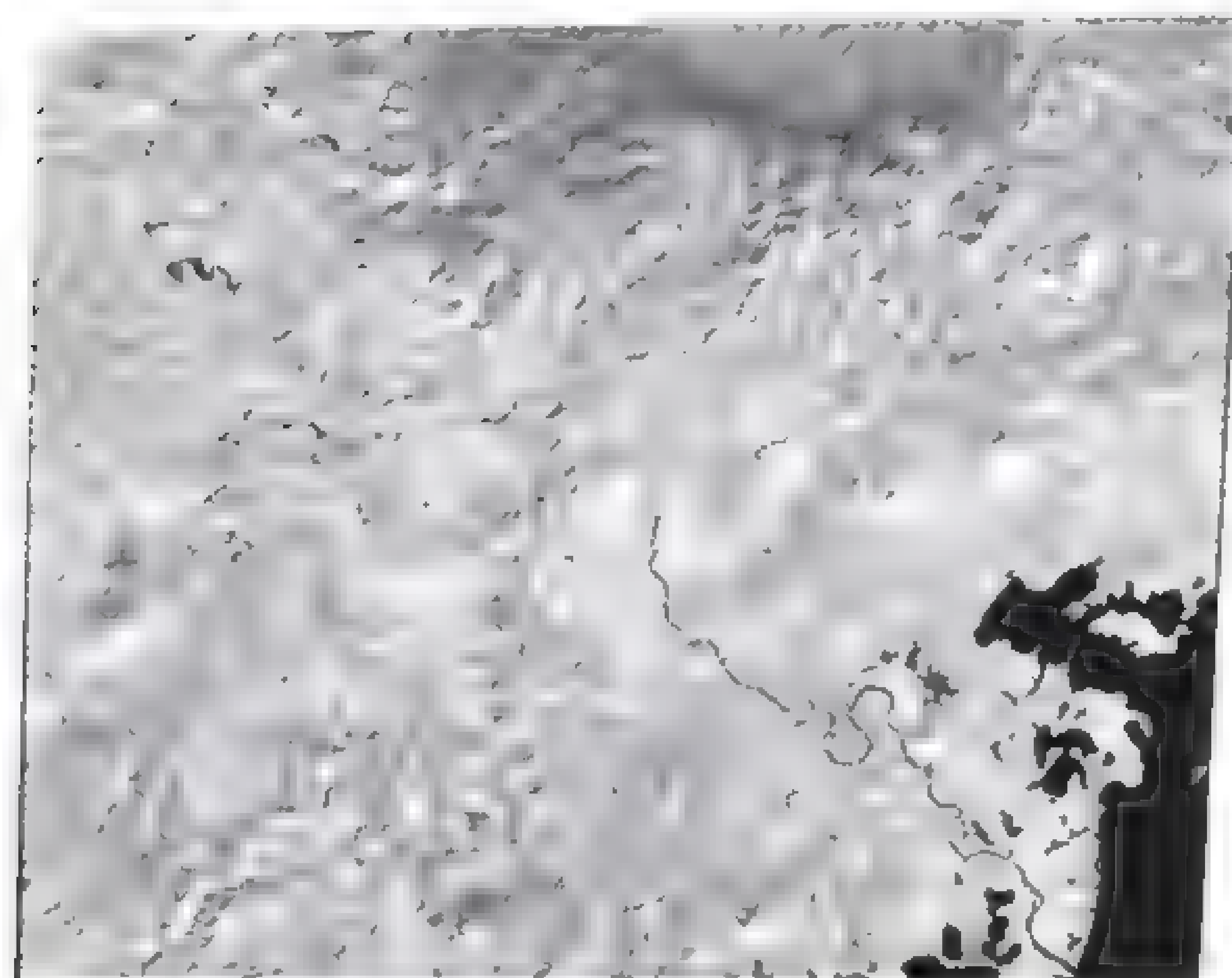


(a) 研究区位置

图3-1 研究对象的地理位置



(b) 位于研究区内的验证区域



(c) 覆盖研究区TM遥感影像(5.4.3波段合成)

图3-1　研究对象的地理位置(续)

图们江地区是东北亚地区的中央区位，具有独特的区位优势，随着图们江地区的开发开放，它的战略地位再度提升。通过有效的分类方法对图们江地区进行土地利用/覆盖信息提取，可为图们江地区土地利用/覆盖动态变化过程分析提供科学的依据，为中、朝、俄土地利用差异的比较提供技术保障[1]。

3.1.2 研究区影像集

本研究选取行列号115-30遥感影像，包括：1992-10-07、2001-08-31、2001-10-02、2006-05-17、2006-09-22、2009-09-30近20年的6幅不同时相的Landsat TM/ETM影像(30米空间分辨率，UTM投影)，数据集详见表3-1。

表3-1 研究区遥感影像数据集

年份	日期	空间分辨率	影像类型
1992	10-07	30m	Landsat TM
2001	08-31 10-02	30m	Landsat ETM+
2006	05-17 09-22	30m	Landsat TM
2009	09-30	30m	Landsat TM

3.1.3 分类体系的建立

土地覆盖类型的确定是土地资源调查和监测的基础，也是开展本项研究的前提[2]~[3]。没有一个理想的土地覆被分类系统适用于所有需求，不同的目标制定不同的分类系统。对于特定的、局部的应用服务，可以制定反映应用目标特征、类型精细的土地覆盖类型，对于大区域或多目标应用服务，可以制定类型定义宽泛的分类系统[4]。本书依据联合国粮农组织提出的土地覆盖分类体系(Land Cover

Classification System，LCCS)和中国科学院资源环境数据库土地利用分类系统，从遥感制图角度和中、朝、俄交界处所处北温带等特点出发，建立了研究区土地覆盖遥感分类体系，详见表3-2，该体系采用二级分类系统，第一级分为6个类型，第二级包括9个类型。

表3-2　研究区土地覆盖遥感分类体系

一级类型	二级类型	含　义
森林	落叶针叶林	郁闭度>30%，高度>2m的落叶针叶天然林和人工林
	落叶阔叶林	郁闭度>30%，高度>2m的落叶阔叶天然林和人工林
	针阔混交林	郁闭度>30%，高度>2m的针阔混交天然林和人工林
	灌丛	郁闭度>40%，高度>2m的灌丛和矮林
草地	典型草地	覆盖度10～30%，以旱生草本为主的草地
农田	旱地	无灌溉水源及设施，靠天然降水生长作物的耕地
建筑用地	居住地	人工硬表面，居住建筑
水体	内陆水体	陆地上各种淡水湖、咸水湖、水库及坑塘、河流
其他	裸地	地表为土质、植被覆盖度在5%以下的裸土地、盐碱地等无植被地段

3.2　遥感影像数字集

3.2.1　样本采集

1. 波段合成

样本的有效采集依赖于遥感影像波段准确的合成。随着遥感技术的不断发展，光谱遥感图像的种类也随之增加。大多数多光谱遥

感图像包含3～10个相对较宽的波段。例如：Aerial Photography、Landsat Multispectral Scanner(MSS)、Landsat Thematic Mapper(TM，ETM+)、SPOT(HRV)、IKONOS与QuickBird[5]。不同波段有不同的用途、波长范围和统计特征，响应不同地物在该波段内的辐射、反射特性。正是由于波段与地物间有这些相关特性，才可以用地物在不同光谱范围的反射程度和波段的组合来识别地物[6]。其中本书采用的TM多光谱影像包括7个波段，如表3-3所示。

表3-3　TM传感器波段特征

波段序号	波谱范围/μm	波段名称	地面分辨率/m	光谱效应
1	0.45～0.52	蓝色	30	对水体有穿透能力，用来分析植被特征、土地利用及编制森林分布图
2	0.52～0.60	绿色	30	对水体的穿透能力较强，对植被的反射敏感，能区分树种、林型
3	0.63～0.69	红色	30	位于叶绿素的吸收区，能增强植被覆盖与无植被覆盖的反差，可判断植被的健康状况
4	0.75～0.96	近红外	30	集中反映植物的强反射，用于生物量、植被类型和作物长势的调查，可绘制水体边界
5	1.55～1.75	中红外	30	处于水的吸收带，对含水量反应敏感，可用于植物含水量、土壤湿度调查和作物长势分析
6	1.04～1.25	热红外	120	对热异常敏感，可监测人类活动的热特征，用于热分布制图、岩石识别和地质探矿
7	2.08～2.35	中红外	30	探测高温辐射源，如监测森林火灾、火山活动等，可区分岩石类型

波段合成通常采用3个波段进行假彩色合成，从TM的7个波段任选3个波段进行组合，加之3个波段所赋基色的不同，其合成方案

总数多达200多种，因此，如何选择最佳组合方案值得深入研究，波段的选取通常考虑3 个方面的因素[7]~[9]：

(1) 波段或波段组合信息含量的多少；

(2) 各波段之间相关性的强弱；

(3) 研究区内欲识别地物的光谱响应特征如何。

信息含量多、相关性小、地物光谱差异大、可分性好的波段就是应该选择的最佳波段。许多学者采用最佳指数因子(OIF)并综合地物的光谱特征进行分析与研究得出：543波段组合为土地利用/覆盖信息提取的最佳波段组合，其合成的假彩色遥感影像具有更好的目视效果，可很好地用于土地利用/土地覆被的目视解译与计算机分类[10]~[12]。因此，本书以TM543波段组合图像进行样本提取。

2. 采样方法

采样方法在某种程度上比分类算法的选择更为重要[13]。为确保样本的代表性、有效性，本书采样方法选择概率采样。常用的概率采样方法包括简单随机采样、分层采样、系统采样以及聚点式集群采样等。具体采用哪种方法，应考虑分类系统和应用目的影响，依据精度评价而定[14]。当样本数量较大时，宜采用简单随机采样方法[15]，而且作为最基本的采样方法，简单随机采样应用广泛，能够满足大多数用户的需求[16]。因此，本书采用简单随机采样方法采集训练样本和测试样本。此外，样本选取的模板窗口大小须合适，基本原则是应该稍大于空间特征单元的大小。如果窗口选取太小，则影响信息提取的精度；而窗口太大，影响训练效率而且并不一定能提高精度；不同大小的模板，能提取不同粗细的特征信息[17]~[18]。

3.2.2 特征选取

用于分类的特征选择旨在提高分类器的泛化能力，降低维数和计算复杂性。它在充分保留分类信息的同时，通过选取特征子集直接降低原始特征维。特征选择已成为分类问题中的研究热点[19]~[20]。通常根据下面的准则选取最小化的特征子集[21]：

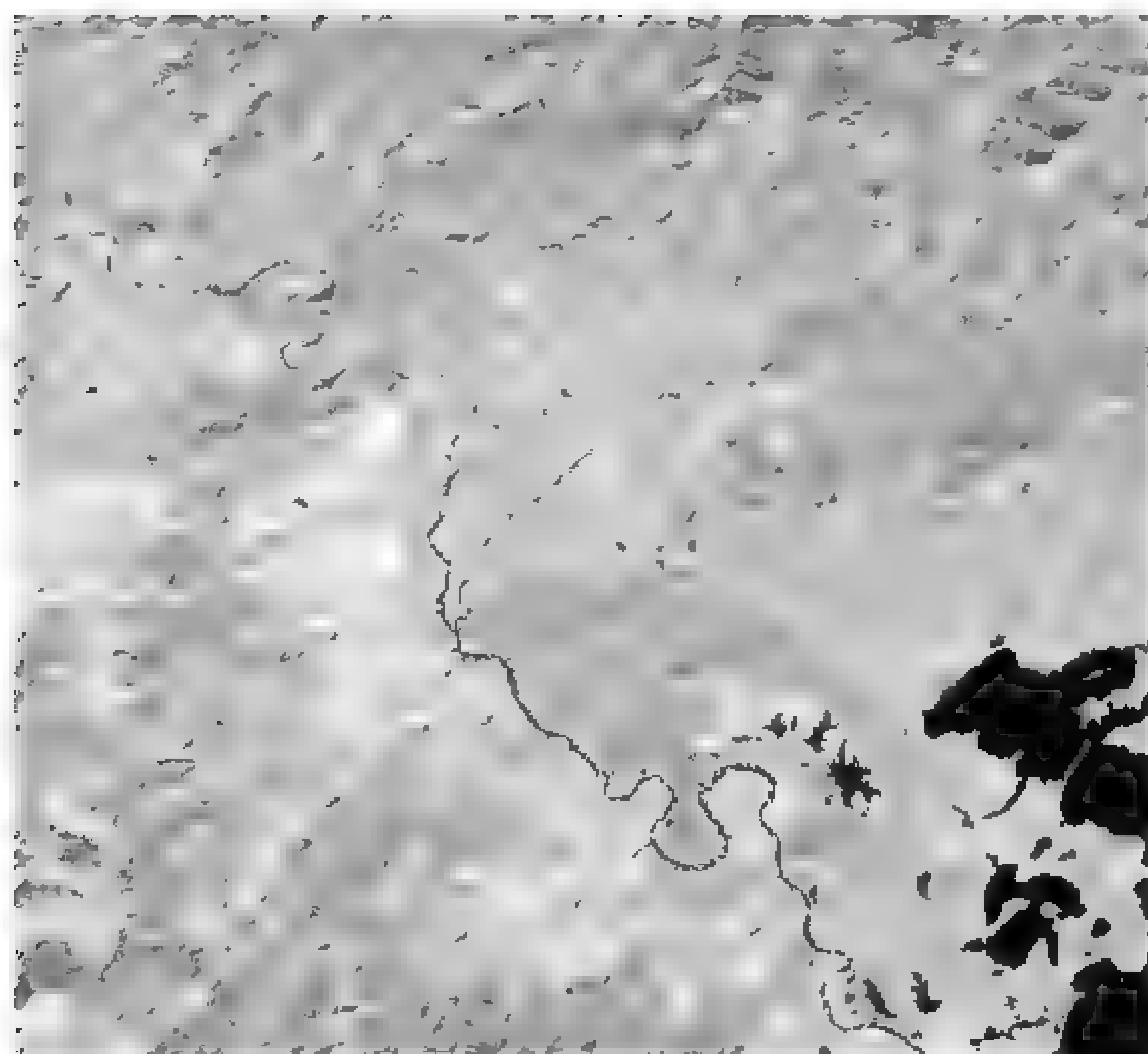
(1) 分类精度不会明显降低；

(2) 特征选取后的类别分布特点与特征选取前原始的类别分布特点尽可能地接近。

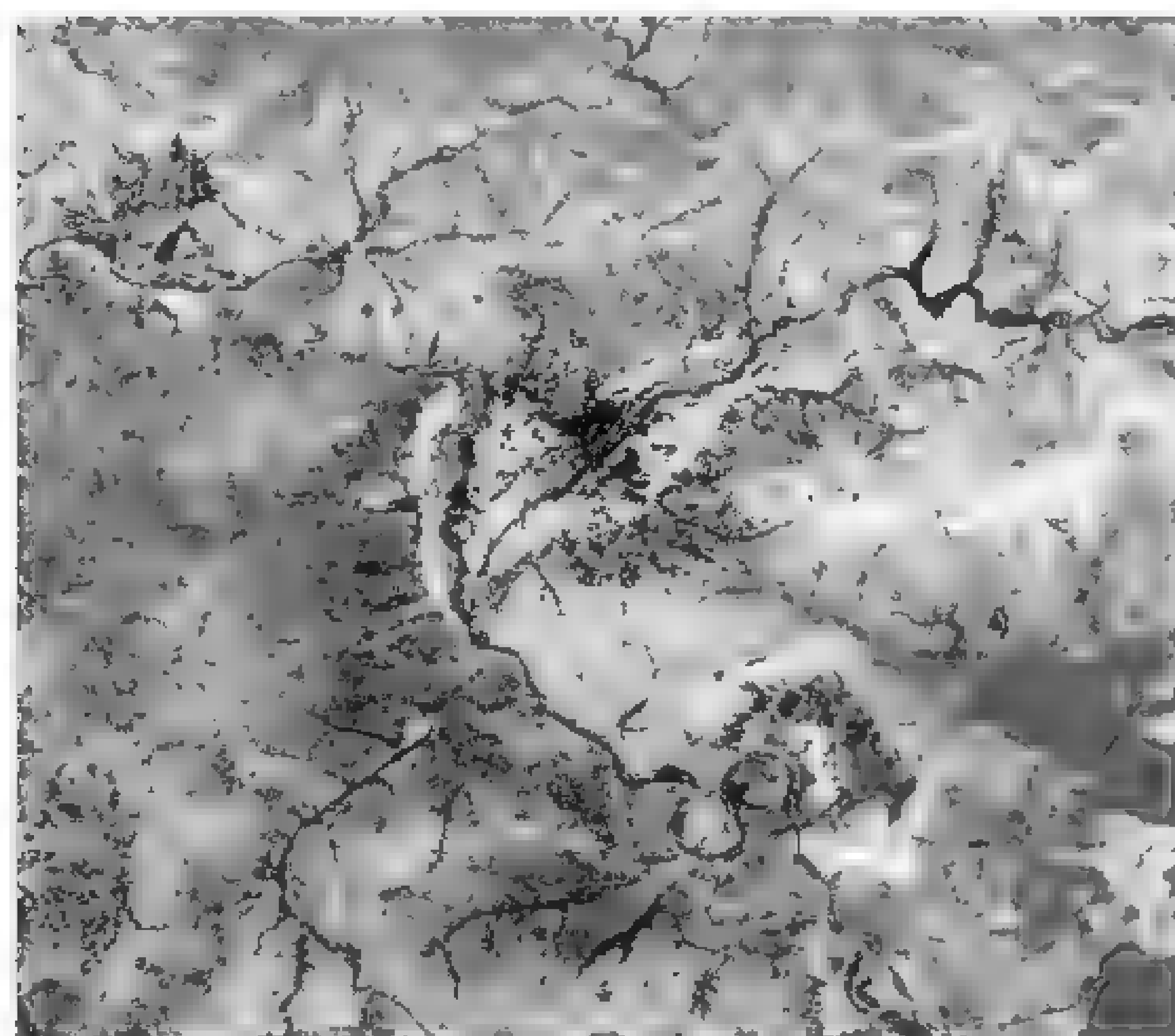
考虑到数据维数及算法复杂度，本文采用8个属性特征，具体包括TM影像的波段信息、PCA主成分分析、NDVI植被指数。

首先对于TM波段选取，由于TM影像的第6波段是热红外波段，其空间分辨率较低(空间分辨率为120m)，一般不参与波段的合成。因此，研究中不把第6波段考虑在内，波段特征包括TM影像的1～5波段、7波段共计6个波段。

遥感图像波段之间存在很强的相关性，从而导致各波段图像之间存在相似的信息和结构特征[22]，造成了冗余信息。主成分分析(PCA)是一种正交分解(Proper Orthogonal Decomposition，POD)算法，等同于Karhunen-Loeve分解(Karhunen-Loeve Decomposition，KLD)和奇异值分解(Singular Value Decomposition，SVD)算法[23]。主成分分析经常被用于多光谱遥感图像的特征提取和图像压缩中[24]~[25]，用于去除多光谱遥感图像各个波段间的重复性和冗余信息。为了更好地比较主成分分析后波段所含信息量的多少，本章以研究区2006-9-22获取的TM影像进行PCA试验，结果如图3-2所示。

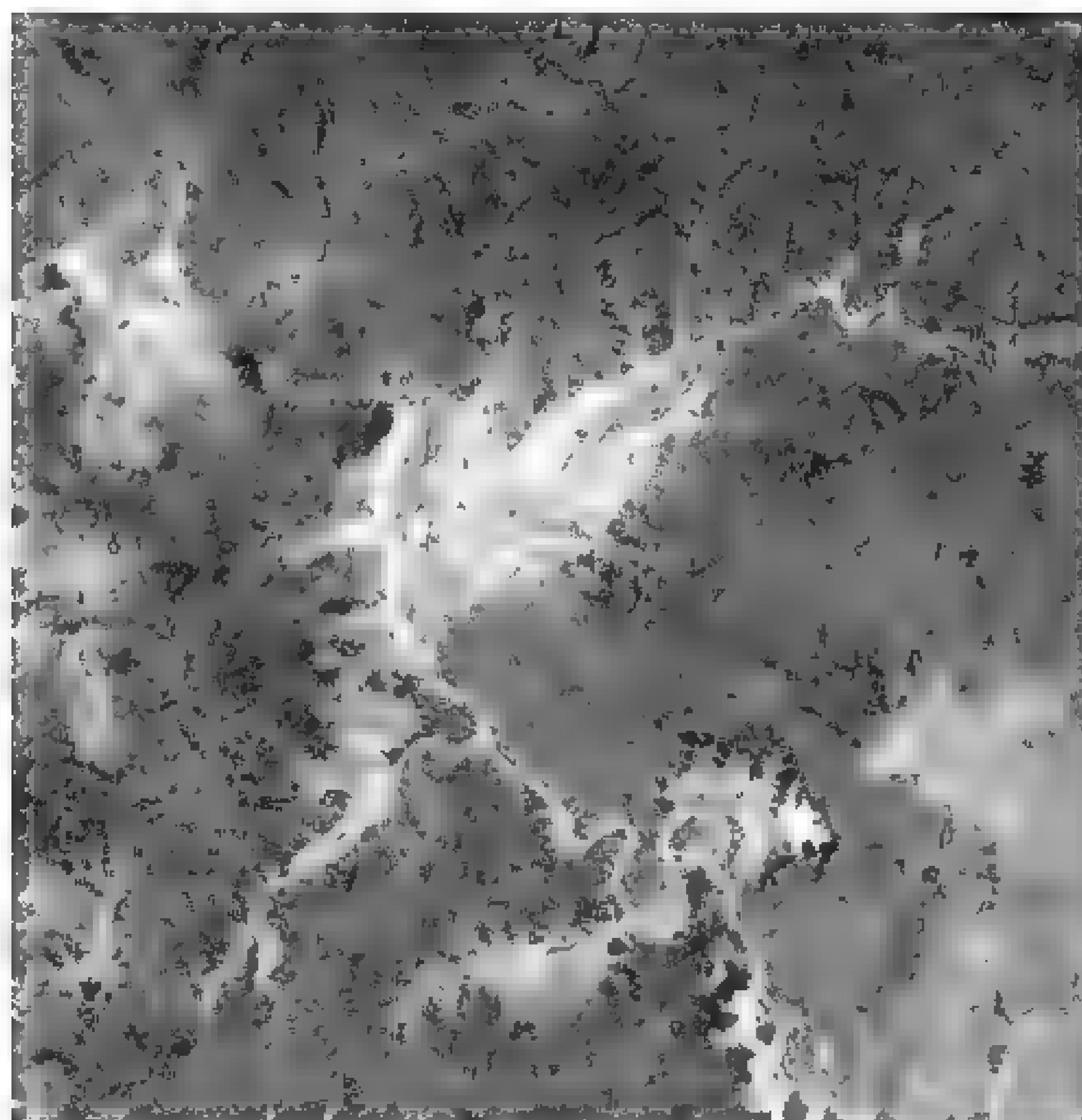


(a) PCA-band 1

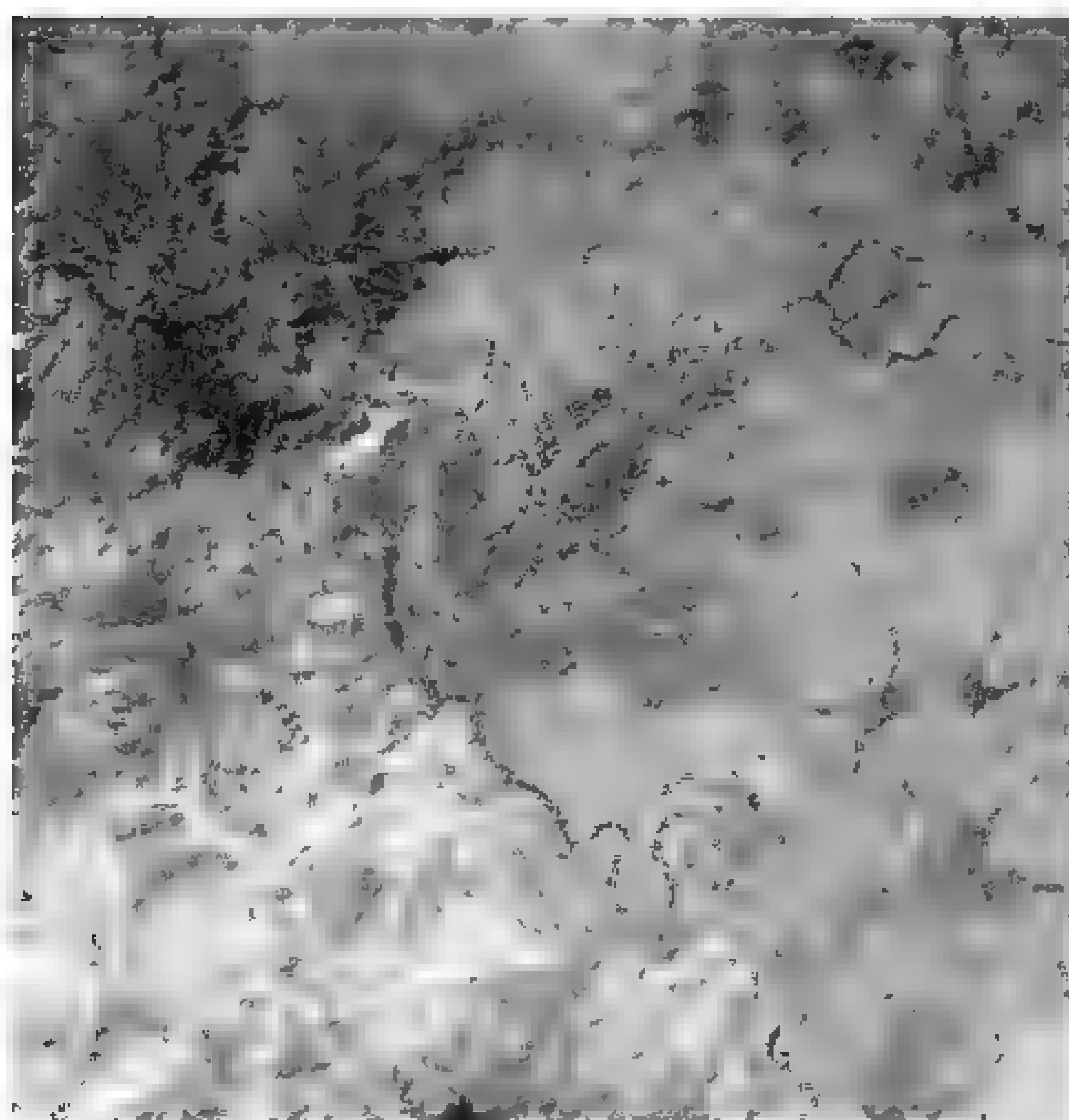


(b) PCA-band 2

图3-2　研究区部分区域PCA主成分分析结果

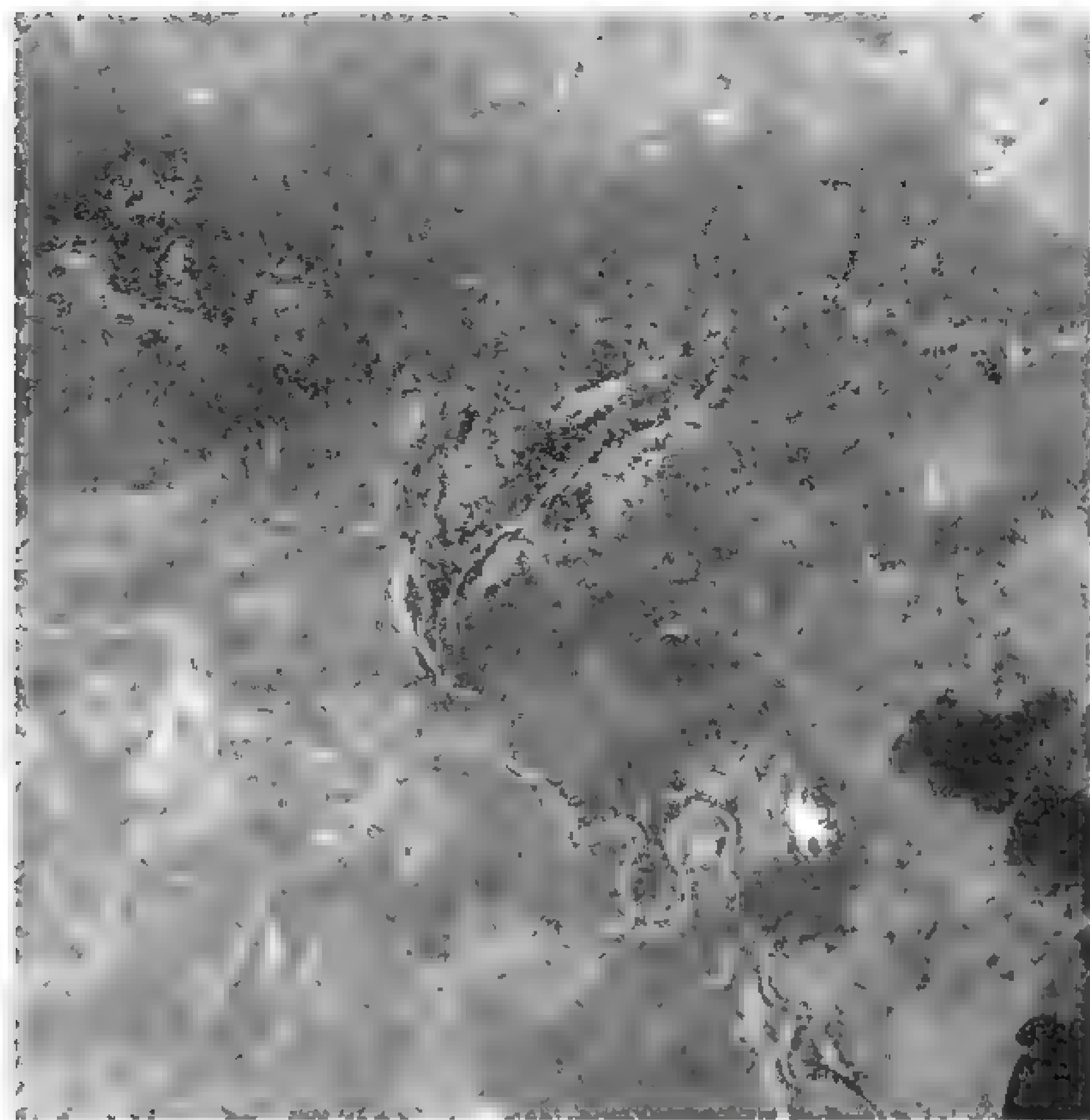


(c) PCA-band 3

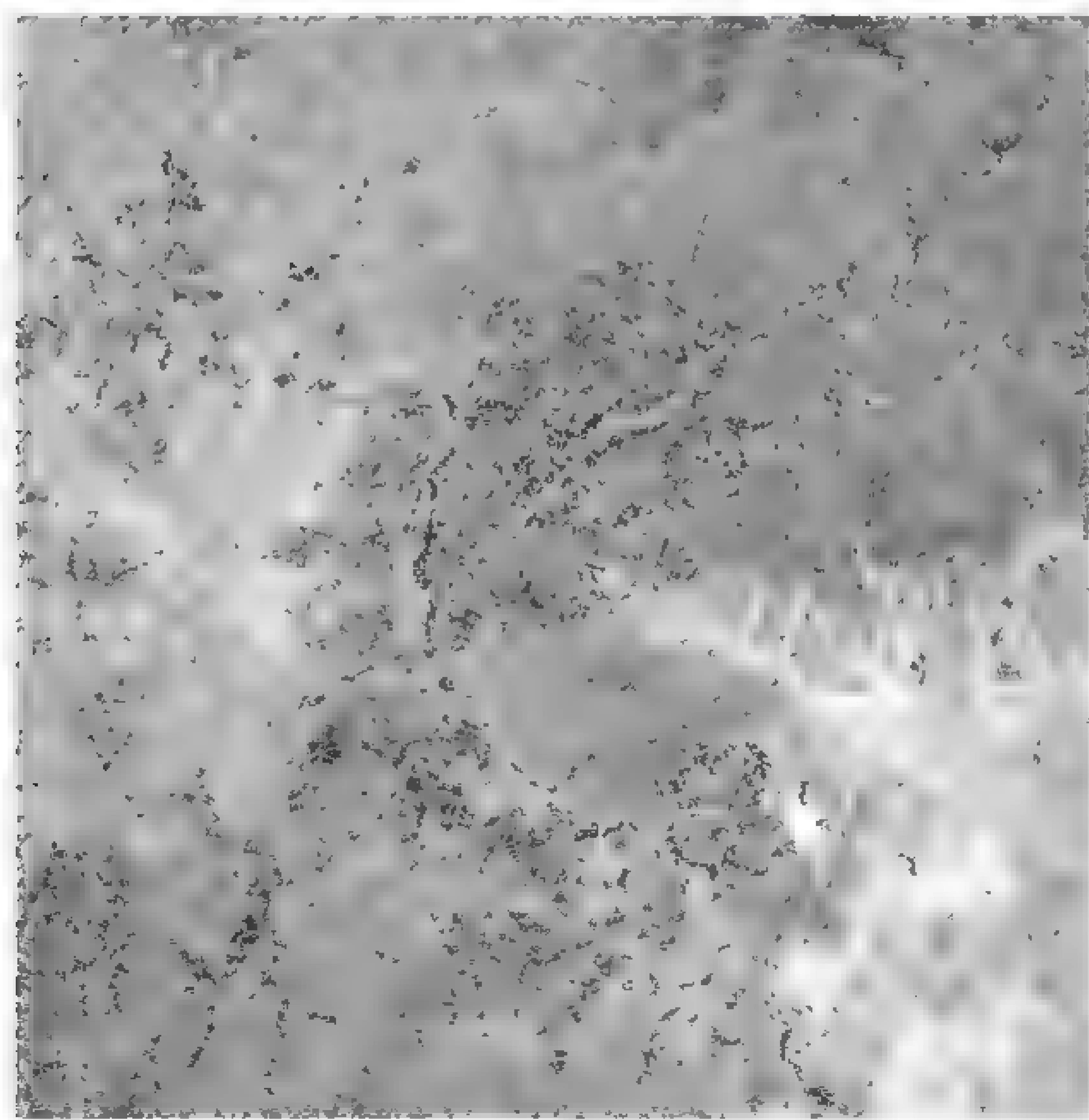


(d) PCA-band 4

图3-2　研究区部分区域PCA主成分分析结果(续)

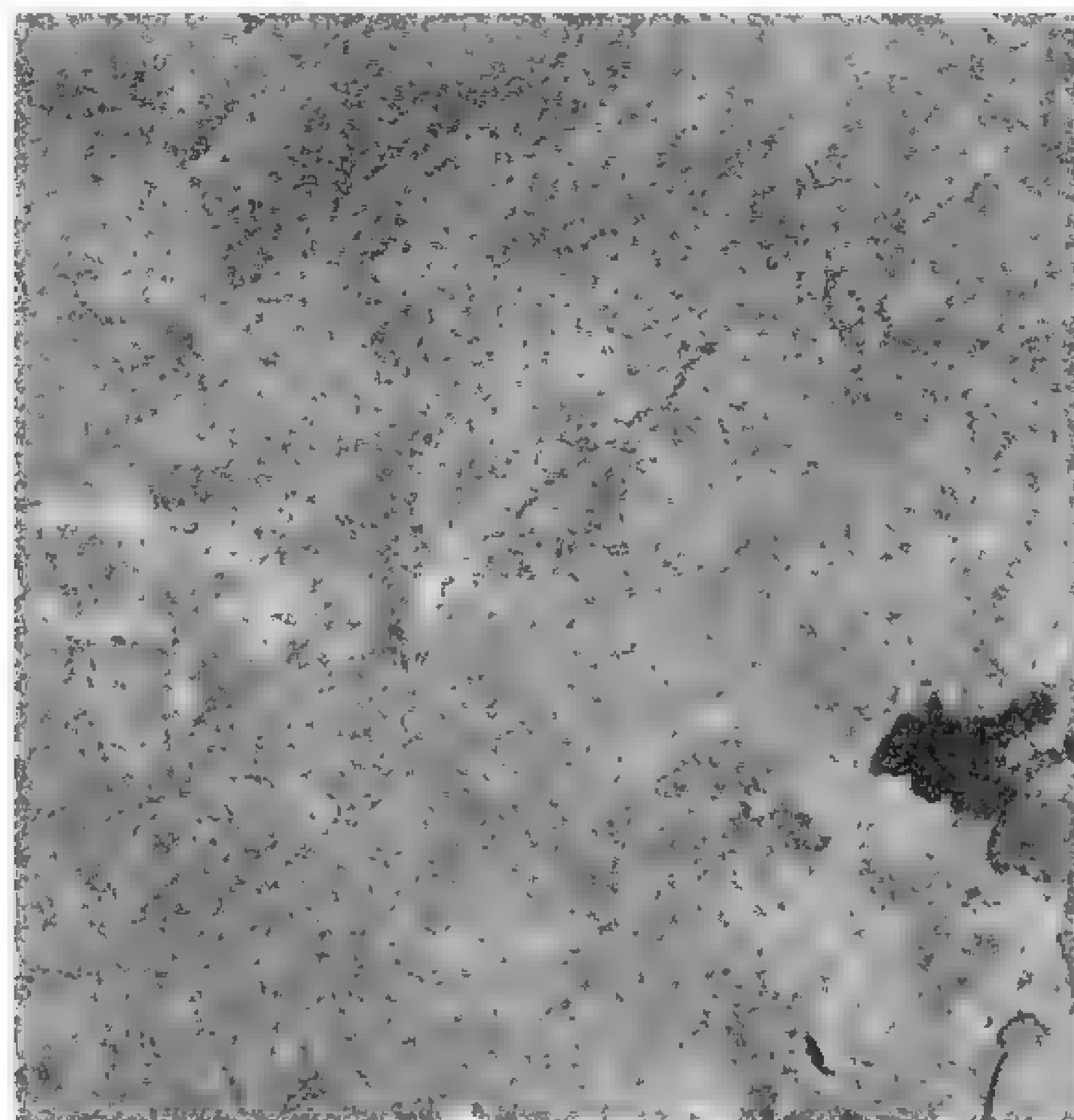


(e) PCA-band 5

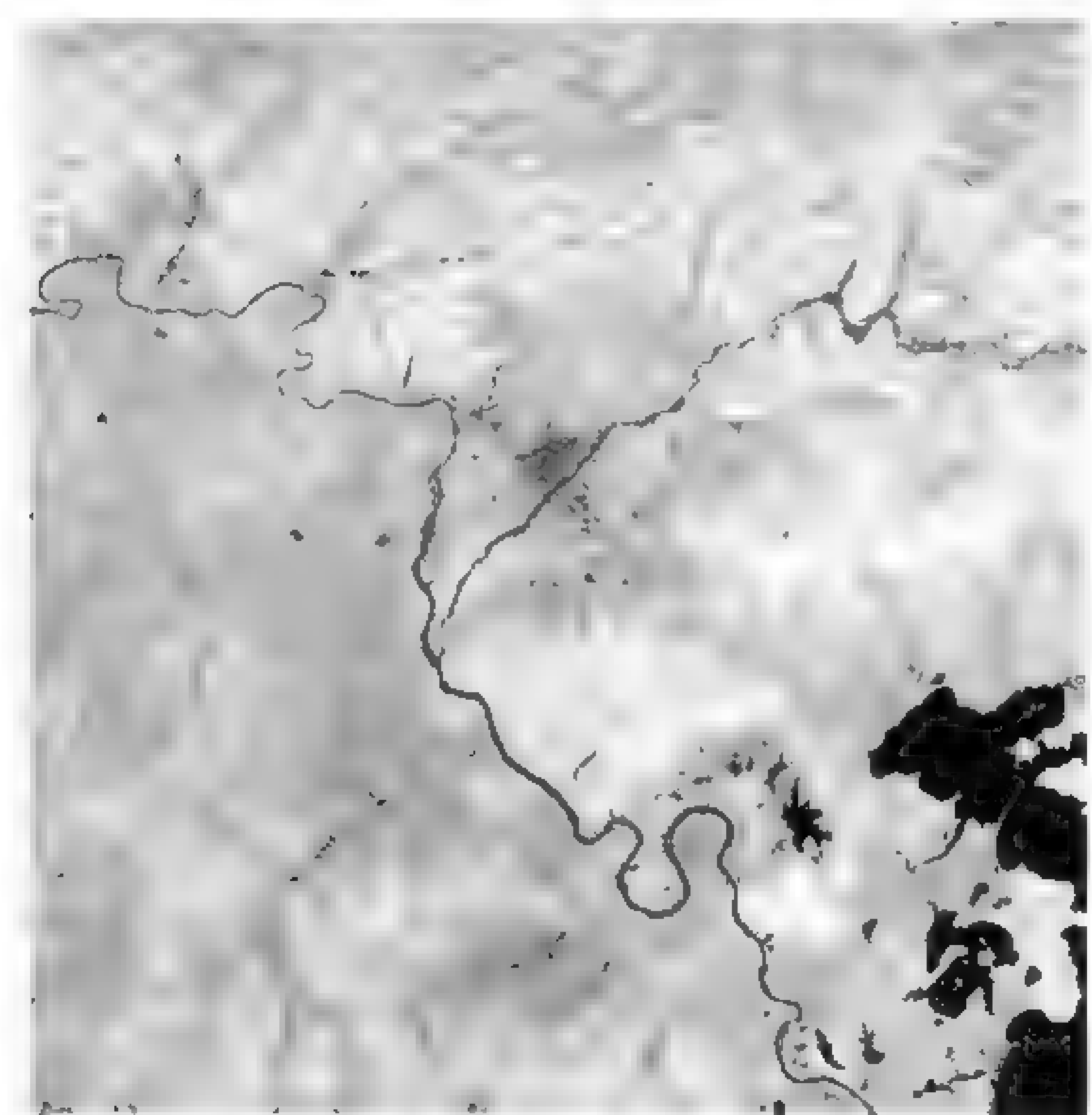


(f) PCA-band 6

图3-2　研究区部分区域PCA主成分分析结果(续)



(g) PCA-band 7



(h) NDVI

图3-2　研究区部分区域PCA主成分分析结果(续)

注：其中(a)～(g)分别为band 1～band 7主成分分析结果，(h)为该研究区植被指数(NDVI)分析结果。

在图3-2(a)中明显能看出band 1所包含信息量最为丰富，相比之下，band 6(图3-2(f))和band 7(图3-2(g))波段信息量很少，基本都是噪声。为了更好的比较，本章也分别列出相应波段PCA特征统计表和特征统计图，如图3-3及表3-4所示。根据图3-3和表3-4也可看出，TM所包含的7个波段中，band 1、band 2和band 3特征值较大，证明包含信息量较多，并且前3个波段里以band 1特征值最大，为493.317 5。因此，本书选择PCA第一主分量作为其中一个分类特征。

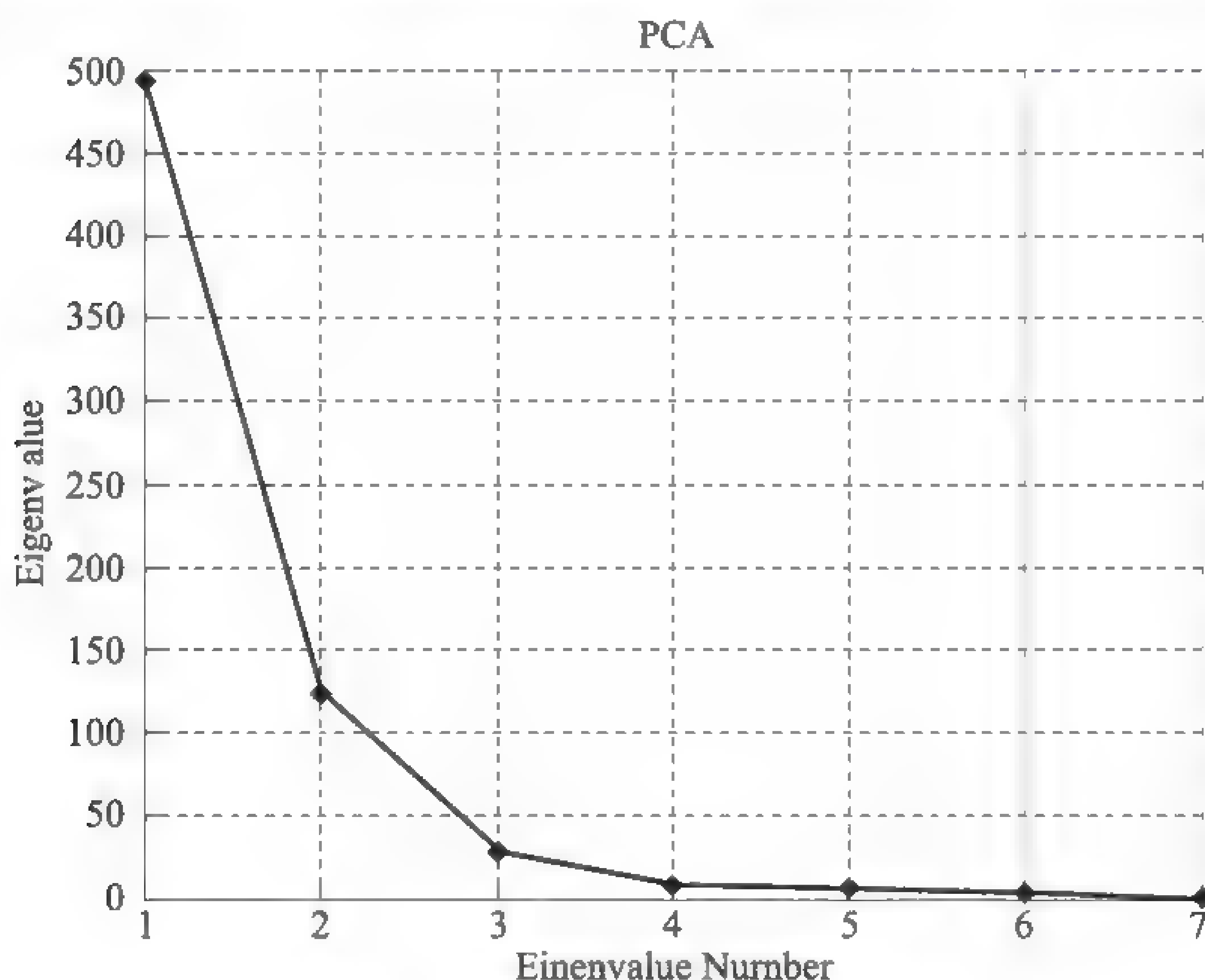


图3-3　PCA主成分分析特征统计表

表3-4　PCA主成分分析特征统计表

PC	Eigenvalue	Percent
Band 1	493.317 5	74.48%
Band 2	123.619 8	93.14%
Band 3	28.704 9	97.47%
Band 4	7.936 5	98.67%

(续表)

PC	Eigenvalue	Percent
Band 5	4.590 3	99.36%
Band 6	3.605 8	99.91%
Band 7	0.604 2	100.00%

不同类型的检索目标，特征应该是不同的，对于某个检索目标内容的描述，不同特征的有效性也是不一样的，如果能够选择最佳表示检索目标内容的特征来进行图像检索，则可以极大地提高检索性能[26]。鉴于研究区的主要土地覆盖类型为植被，因此最后采用的特征为归一化植被指数。归一化植被指数(Normalized Difference Vegetation Index，NDVI)是评估植被状况的最有效参数之一[27]，如公式3-1所示。NDVI能很好地反映植被生长状态和植被覆盖度，也是衡量生态系统的重要信息，试验结果如图3-2(f)所示，从图中不难看出研究区所含植被信息丰富。

$$NDVI = \frac{(B_4 - B_3)}{(B_4 + B_3)} \tag{3-1}$$

3.3　本章小结

本书选择图们江下游，中、朝、俄交界处作为研究对象。以行列号115-30一景、近20年的6幅不同时相的Landsat ETM/TM影像作为研究材料，分别讨论本书所采用的影像合成方式、特征采集方法、土地覆盖分类依据，以及特征选取方法，为进一步研究分类方法提供必要的数据材料。

参考文献

[1] 南颖, 吉喆, 董叶辉, 倪晓娇. 30年来图们江跨国界地区土地利用/覆盖动态变化研究[J]. 湖南师范大学(自然科学版), 2012, 35(1): 82-89.

[2] 张增祥, 汪潇, 王长耀. 基于框架数据控制的全国土地覆盖遥感制图研究[J]. 地球信息科学学报, 2009, 11(2): 216-223.

[3] 刘勇洪, 牛铮, 徐永明. 基于MODIS数据设计的中国土地覆盖分类系统与应用研究[J]. 农业工程学报, 2006, 22(5): 99-104.

[4] 吴炳方, 张磊, 李晓松. 面向生态系统碳收支服务的全国土地覆被分类系统设计方案[J]. 自然资源学报, 2011.

[5] 杨晨. 基于机器学习的土地覆盖遥感信息提取方法研究[R]. 吉林大学, 2010.

[6] 卢玉东, 尹黎明, 何丙辉, 宋光煜, 熊有胜. 利用TM影像在土地利用 覆盖遥感解译中波段选取研究[J]. 西南农业大学学报, 2005, 4 (27): 479-486.

[7] 苏红军, 杜培军, 盛业华. 一种基于分形维数的高光谱遥感波段选择算法研究[J]. 测绘通报, 2007, 3: 23-26.

[8] 许菡, 燕琴, 徐泮林. 多源遥感影像融合最佳波段选择及质量评价研究[J]. 测绘科学, 2007, 32(3) : 72-87.

[9] 徐磊, 侯立春, 杨强, 张志, 金姝兰. 利用TM影像提取土地利用/覆被信息的最佳波段研究[J]. 湖北大学学报(自然科学版), 2011, 1(33): 119-122.

[10] Tuia D., Pacifici F., Kanevski M., Emery W.J.. *Classification of*

Very High Spatial Resolution Imagery Using Mathematical Morphology and Support Vector Machines[J]. IEEE Transactions on Geoscience and Remote Sensing, 47(11): 3866-3879.

[11] 张韬, 吕洪娟, 孙美霞. 遥感多光谱数据在内蒙古西部湿地监测中最佳波段选取的应用研究[J]. 干旱区资源与环境, 2007, 21(4): 102- 106.

[12] 周旭, 安裕伦, 张斌. CBERS-CCD 数据土地利用/覆盖信息提取最佳波段选择——以贵州喀斯特山区为例[J]. 遥感技术与应用, 2009, 24(6): 743-749.

[13] Hellden, U.. *A Test of Landsat-2 Imagery and Digital Data for Thematic Mapping Illustrated by an Environmental Study in Northern Kenya*[R]. Sweden: Lund University Natural Geography Institute Report, 1980: 47.

[14] 赵英时. 遥感应用分析原理与方法[M]. 北京:科学出版社, 2003.

[15] Giles, M. F.. *Status of Land Cover Classification Accuracy Assessment*[J]. Remote Sensing of Environment, 2002, 80: 185-201.

[16] Stehman S.V., Czaplewski R.L. *Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles*[J]. Remote Sensing of Environment, 1998, 64: 331-344.

[17] 李建平. 基于FNMPSVM模型和图像分割的盐碱地信息提取研究[R]. 中国科学院, 2007.

[18] 骆剑承, 周成虎, 梁怡. 支撑向量机及其遥感影像空间特征提取和分类的应用研究[J]. 遥感学报, 2002, 6(1): 50-55.

[19] Fröhlich H., Chapelle O.. *Feature Selection for Support*

Vector Machines by Means of Genetic Algorithms[C]. The 15th IEEE International Conference on Tools with Artificial Intelligence. USA, Sacramento, CA, 2003: 142-148.

[20] Liu Y., Zheng Y.F.. FS SFS: *A Novel Feature Selection Method for Support Vector Machines*[J]. Pattern Recognition, 2006, 39:1333-1345.

[21] Dash M., Liu H.. *Feature Selection for Classification*[J]. Intelligent Data Analysis, 1997, 1:131-156.

[22] Bryant J.. *On Displaying Multispectral Imagery*[J]. Photogrammetric Engineering and Remote Sensing, 1988, 54(12): 1739-1743.

[23] Liang Y.C., Lee H.P., Lim S.P., Lin W.Z., Lee K.H., Wu C.G.. *Proper Orthogonal Decomposition and Its Applications-Part I:Theory*[J]. Journal of Sound and Vibration, 2002, 252(3): 527-544.

[24] Zhao G., Maclean A.L.. *A Comparison of Canonical Discriminant Analysis and Principal Component Analysis for Spectral Transformation*[J]. Photogrammetric Engineering & Remote Sensing, 2000, 66(7): 841-847.

[25] Mitternicht G.I., Zinck J.A.. *Remote Sensing of Soil Salinity: Potentials and Constraints*[J]. Remote Sensing of Environment, 2003, 85: 1-20.

[26] 朱佳丽, 李士进, 万定生, 冯钧. 基于特征选择和半监督学习的遥感图像检索[J]. 中国图象图形学报, 2011, 16(8): 1474-1482.

[27] Paruelo J.M., Epstein H. E., Lauenroth W. K.. *Anpp Estimates from NDVI for the Central Grassland Region of the United States*[J]. Ecology, 1997, 78(3): 953-957.

第4章

SVM参数优化方法研究

4.1 SVM理论及参数优化算法研究进展

4.1.1 SVM的核心思想

SVM的核心思想是把样本通过非线性映射投影到高维特征空间，以结构风险最小化原理(Structural Risk Minimization，SRM)为原则，在高维特征空间中构造VC维(Vapnik-Chervonenkis Dimension，即描述函数集或学习机器的复杂性)，以尽可能低的最优分类超平面作为分类面，使分类风险上界最小化，从而使分类算法具有最优的推广能力。其中核函数及其参数选择是SVM面对的重要而困难的问题，对提高遥感影像分类精度是非常重要的。

4.1.2 SVM理论

Vapnik[1]所提出SVM理论是在20世纪90年代中期不断发展和成熟的。SVM是一种新的、有效的统计学习方法，是近年来模式识别与机器学习领域的一个新的研究热点。

SVM的主要思想是建立一个超平面作为决策曲面，使得正例和反例之间的隔离边缘(Margin Width)被最大化。如图4-1所示的两维情况，图中实心点和空心点代表两类样本，虚线为分类线，两条实线分别为各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔(Margin)。所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为0)，而且要使分类间隔最大。

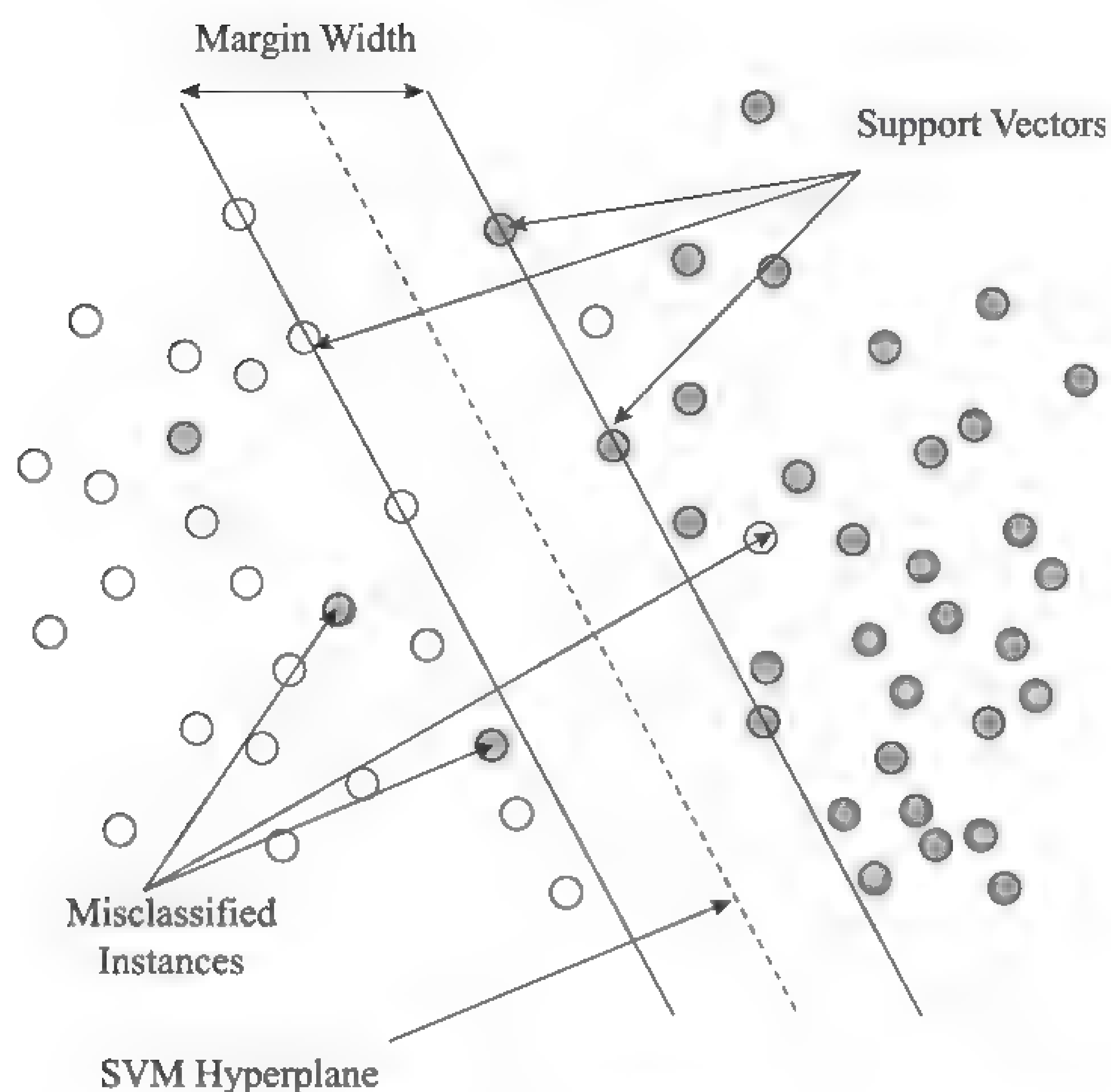


图4-1　SVM理论描述

SVM不仅考虑了对渐进性能的要求，而且追求在有限信息的条件下得到最优结果，避免了人工神经网络等方法的网络结构选择、过学习和欠学习以及局部极小等问题。其线性可分模式的最优超平面的推导如下：

考虑训练样本$\{x_i, y_i\}_{i=1}^{N}$，其中x_i是输入模式的第i个样本，$y_i \in \{-1, +1\}$。

设用于分离的超平面方程是：

$$\omega \cdot x + b = 0 \tag{4-1}$$

其中ω是可调的权值向量，b是偏置，在此设最优的ω和b为ω_0和b_0，则最优的分类超平面为：

$$\omega_0 x + b_0 \quad 0 \tag{4-2}$$

满足下面条件的特殊数据点称为支持向量$\{x_i, y_i\}$，且满足：

$$\omega \cdot x_i+b=1,\ y_i=1 \tag{4-3}$$

或者

$$\omega \cdot x_i+b=-1,\ y_i=-1 \tag{4-4}$$

此处设x_1和x_2为两个支持向量，且满足：

$$\omega \cdot x_1+b=1 \tag{4-5}$$

$$\omega \cdot x_2+b=-1 \tag{4-6}$$

公式(4-5)与公式(4-6)相减可得：

$$\omega \cdot (x_1-x_2)=2 \tag{4-7}$$

进而可得Margin Width=$\frac{\omega}{\|\omega\|}\cdot(x_1-x_2)=\frac{2}{\|\omega\|}$，则$\frac{2}{\|\omega\|}$最大化⇔$\|\omega\|$最小化⇔$\frac{\|\omega\|^2}{2}$最小化。

且对于任意的(x_i, y_i)有：

$$\begin{cases}\omega \cdot x_i+b \leqslant -1,\ y_i=-1\\ \omega \cdot x_i+b \geqslant 1,\ y_i=1\end{cases} \tag{4-8}$$

可得：

$$y_i(\omega \cdot x_i+b) \geqslant 1 \tag{4-9}$$

寻找最优超平面即正反例间隔最大化问题，最终归结为一个二次规划问题，可使用Lagrange乘子法解决。

首先建立lagrange函数：

$$J(\omega,b,a)=\frac{1}{2}\omega^T\omega-\sum_{i=1}^{N}a_i\left[y_i(\omega \cdot x_i+b)-1\right] \tag{4-10}$$

其中a_i称为Langrange乘子。对ω，b求偏导并置零，有：

$$\frac{\partial J}{\partial \omega}=0 \Leftrightarrow \omega=\sum_{i=1}^{N}a_i y_i x_i \tag{4-11}$$

$$\frac{\partial J}{\partial \omega}=0 \Leftrightarrow \sum_{i=1}^{N}a_i y_i=0 \tag{4-12}$$

最终可以得到原问题的对偶问题：

$$\max_a Q(a) = J(\omega,b,a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j x_i^T x_j \qquad (4\text{-}13)$$

$$\sum_{i=1}^{N} a_i y_i = 0, a_i \geqslant 0 \qquad (4\text{-}14)$$

若数据非线性可分，可以引入一个非线性映射φ：$x_i \to \varphi(x_i)$将其训练数据映射到高维特征空间中，使之在这个高维空间中线性可分，但由于没有数据的先验知识，这个非线性映射是很难知道的。SVM的一个很重要的特点就是引入一个核函数$K(x_i, x)$来代替高维空间中的内积，如图4-2所示，将样本点从低维空间映射到高维Hilbert空间，在高维特征空间中设计线性最优分离超平面，得到输入空间中的非线性学习算法，使低维非线性问题转化为高维线性可分问题。引入核函数后，公式(4-13)可表示为：

$$\max_a Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \qquad (4\text{-}15)$$

设最优Lagrange乘子为a^*，则最终的最优判别函数为：

$$f(x) = sign(\sum_{i=1}^{N} a_i^* y_i K(x_i x) + b) \qquad (4\text{-}16)$$

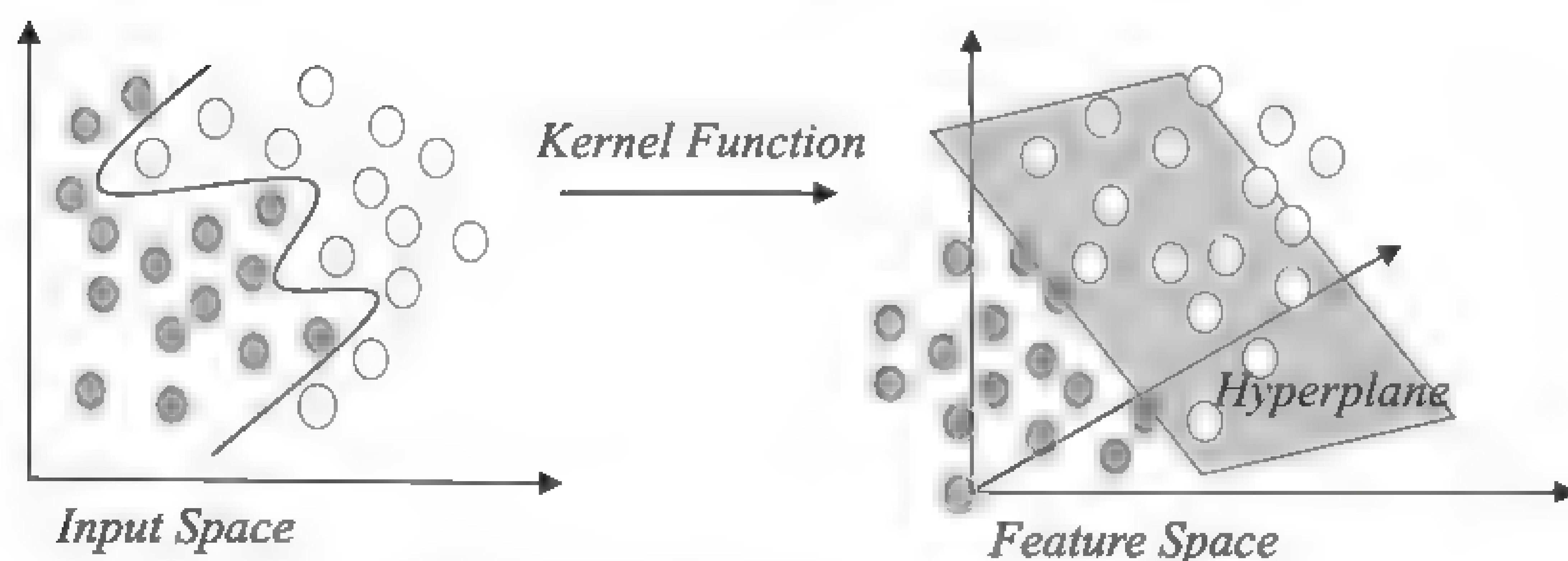


图4-2　核函数参数实现数据集的高维空间非线性映射

下面列出几种常用的核函数：

(1) 线性核函数：$K(x_i, x)=(x_i \cdot x)$ (4-17)

(2) 多项式核函数：$K(x_i, x)=(x_i \cdot x+1)^d$，其中$d$为自然数 (4-18)

(3) 径向基核函数：$K(x_i,x)=\exp(-\frac{1}{\sigma^2}\|x_i-x\|^2)$ (4-19)

或者$k(x_i, x)=\exp(-\gamma\|x_i-x\|)$ (4-20)

(4) Sigmoid核函数：$K(x, x_i)=\tanh(v(x \cdot x_i)+c)$ (4-21)

4.1.3 SVM参数优化方法研究进展

SVM分类属于监督分类，给定训练样本及所对应的类别，通过某种算法从样本计算出分类模型并将该模型推广到待分类点，根据判别函数的指示，最终得出该分类点所属的类别。SVM结构简单，泛化能力强，容易解决具有高维特征、小样本与不确定性等问题，已广泛应用在遥感土地覆盖分类领域[2]~[5]。

SVM的分类参数主要指惩罚参数c和核函数参数。惩罚系数c是一个重要的量，控制对错分样本的惩罚程度。通常，c值越大，两个超平面之间的最大间距越小，错分样本越少，训练时间越长；反之，两个超平面之间的最大间距越大，错分样本越多，训练时间越短。每个数据子空间至少存在一个合适的c，使得SVM推广能力最好。当c超过一定值时，SVM的复杂度达到了数据子空间允许的最大值，此时经验风险和推广能力几乎不再变化。然而，目前还没有一个统一的方法来决定c的最佳取值。而核参数的改变实际上是隐含地改变映射函数，从而改变样本数据子空间分布的复杂程度，即线性分类面的最大VC维。核函数取值过小，所有的样本都被视为支持向量，故而造成对新样本的测试时间长，并且会产生“过度拟合”现象；而当核函数很大时，SVM的性能也会非常差，它对新样

本的正确分类能力几乎为零，将把所有样本都判为同一类。因此，SVM参数的正确选择对分类器泛化能力有着重要的影响[6]。通常，SVM参数选择比较常用的方法包括穷举法、网格法和智能优化法。

1. 穷举法

该方法是在模型选择以后，首先为常数c和核函数固有的参数赋初始值，然后开始实验测试，根据测试精度重复调整参数值，直至得到满意的测试精度为止。王睿[7]在分析SVM原理的基础上，分析了SVM中核函数、核参数及惩罚参数c对分类精度的影响，利用试凑法、最优化法两种方法对SVM参数进行优化比较。

穷举法基本是凭经验调整参数值，缺乏足够的理论依据，对不同的核函数、不同的样本，其调整方法可能不同，因此，在参数调整过程中带有一定的盲目性，且当需要较大幅度调整时，调整次数较多，实验比较复杂。

2. 网格法

网格搜索法是将c和核函数(如径向基核函数参数γ)分别取M个值和N个值，用M×N个(c，γ)的组合分别训练不同的SVM，再估计其学习精度，进而在M×N个(c，γ)的组合中得到学习精度最高的一个组合作为最优参数。吴渝等[8]针对SVM分类面过于复杂和过学习现象，提出基于网格的最近邻SVM算法。李京华等[9]针对网格搜索SVM参数的方法存在复杂度高、运算量大等不足，提出了一种改进的网格搜索SVM分类器的最佳参数选择算法。Hsu和Lin[10]利用工作集的正确选择以及网格搜索SVM分类参数提高分类模型的泛化能力。

然而，网格搜索法虽然可以并行处理多个SVM的训练，但其计

算量为$O(N^2)$，因此，完成一个完全的网格搜索是非常费时的，结果不是很理想[11]。

3. 智能优化法

鉴于以上两种方法的缺陷，一些学者更倾向于利用智能算法来选取SVM参数，其中以遗传算法(Genetic Algorithms，GA)和粒子群算法(Particle Swarm Optimization，PSO)被采用的频次居多。Fröhlich和Chapelle[12]采用GA实现SVM特征子集的选择与参数同步优化，并取得了很好的效果。Zheng和Jiao[13]利用GA对SVM参数自动选取。虽然GA相对于网格法和穷举法在计算时间上降低了不少，但是该算法对进化的每一代种群个体都需要编码和译码，实现过程比较烦琐，同时它也缺乏有效的局部搜索机制，在接近全局极值时收敛速度往往过慢，获取最优个体的时间代价依然较大。因此，也有很多学者利用PSO对SVM参数优化，Vahid等[14]在解决控制图模式识别问题中采用PSO实现SVM分类器优化模型构建。Huang和Dun[15]利用离散PSO和连续值PSO同时优化输入特征子集，并对SVM核函数参数进行选择，将混合PSO-SVM分布式运行以降低时间复杂度。丁胜等[16]采用PSO算法自动选择合适的波段影像并对SVM核函数参数进行优化，提出一种新的PSO-BSSVM分类模型，对高光谱遥感影像的分类试验，比较证明该模型分类性能更优。尽管在寻优过程中PSO相比GA方法，没有GA算法的选择、交差、变异过程，算法收敛速度快，算法结构简单。但PSO也同时存在着容易早熟收敛、搜索精度较低、后期迭代效率不高等缺点。

4.2　基于自适应变异粒子群参数优化的土地覆盖分类模型

专著针对传统PSO优化SVM参数存在早熟收敛、后期迭代效率不高而造成参数寻优不准确的缺点，探索了一种基于自适应变异粒子群算法(Adaptive Mutation Particle Swarm Optimization，AMPSO)的SVM参数优化模型(AMPSO-SVM)。自适应粒子群优化算法的主要思想是以粒子群的群体适应度方差σ^2和全局极值与理论最优值的比较作为粒子群优化算法是否陷入局部极值的评价指标，引入变异算子使得算法能够及时跳出局部极值进而获得全局最优解，接下来介绍详细算法。

4.2.1　传统粒子群算法(PSO)

粒子群优化(Particle Swarm Optimization，PSO)由Kennedy和Eberhart等人提出的模拟群体(Swarm)智能行为的优化算法(Kennedy，1995)。PSO源于对鸟群捕食的行为模拟，每只鸟被称为一个粒子，每个粒子用其几何位置和速度向量表示[17]。PSO的基本原理是将系统初始化为一组随机解，通过迭代搜寻最优值。在每一轮的迭代中，粒子通过速度更新当前位置，并通过适应度函数计算出其适应值，最后根据以下公式，更新粒子的当前速度和位置。

$$V_{id}^{k+1}=\omega V_{id}^{k}+c_1r_1(P_{id}^{k}-X_{id}^{k})+c_2r_2(P_{gd}^{k}-X_{id}^{k}) \qquad (4\text{-}22)$$

$$X_{id}^{k+1}=X_{id}^{k}+V_{id}^{k+1} \qquad (4\text{-}23)$$

其中，ω是惯性系数，为非负数；粒子集$X_i=(X_{i1}，X_{i2}，\cdots，X_{id})$，它在空间的飞行速度用$V_i=(V_{i1}，V_{i2}，\cdots，V_{id})$表示；$d=1，2，$

3，…，D；c_1和c_2是两个正常数，称为学习因子；rand表示[0，1]的随机数。公式(4-22)表示粒子在第$(t+1)$次迭代时的速度由第t次迭代时的速度和局部最优及全局最优粒子位置共同决定，其中，P_i（P_{i1}，P_{i2}，…，P_{id})表示粒子i的局部最优值，即粒子i到目前为止在搜索空间中的最佳点；P_g（P_{g1}，P_{g2}，…，P_{gd})表示整个粒子群的全局最优值，即整个粒子群到目前为止在搜索空间中的最佳点；公式(4-23)是粒子的位置更新公式，表示粒子在第$(t+1)$次迭代的位置等于粒子在第t次迭代时的位置和速度之和。迭代终止条件根据具体问题一般选最大迭代次数或最优位置满足预定最小适应阈值。

4.2.2　自适应变异粒子群优化算法(AMPSO)

粒子群优化算法多采用实数编码，没有遗传算法选择、交差、变异过程，算法收敛速度快，算法结构简单，如果某个粒子发现一个当前的最优位置，其他粒子将迅速向其靠拢，假设该位置非全局最优点，仅为该位置的局部最优点，粒子将无法在解空间内重新搜索，进而出现算法产生局部最优解，即所谓早熟收敛现象。为了克服早熟收敛，就必须提供一种机制，当PSO发生早熟收敛之时，能够跳出局部极值，进入到解空间的其他区域继续搜索，直到最后找到全局极值。吕振肃和侯志荣[18]借鉴遗传算法中的变异思想，在PSO算法中引入变异操作，提出自适应变异粒子群算法，具体描述如下。

粒子群优化算法无论是早熟收敛还是全局收敛，粒子群中的粒子都会出现“聚集”现象。也许所有粒子聚集在某一特定位置，也许聚集在某几个特定位置，这主要取决于适应度函数的选择以及问

题本身的特性。通常粒子位置的一致等价于各粒子的适应度相同。因此，研究粒子群中所有粒子适应度的整体变化就可以跟踪粒子群的状态。为了定量描述粒子群的状态，首先给出群体适应度方差的定义，同时也给出了粒子收敛的定义。

定义1　设粒子群的粒子数目为n，f_i为第i个粒子的适应度，f_{avg}为粒子群目前的平均适应度，σ^2为粒子群的群体适应度方差，则σ^2可以定义为：

$$\sigma^2=\sum_{i=1}^{n}\left[\frac{f_i-f_{avg}}{f}\right]^2 \tag{4-24}$$

其中f是归一化定标因子，其作用是限制σ^2的大小，f可以取任意值，满足两个条件：①归一化后，整个粒子群$|f-f_{avg}|$的最大值不大于1；②f随算法的进化而变化，f的取值采用如下公式：

$$f=\begin{cases}\max\{|f_i-f_{avg}|\}, & \max\{|f_i-f_{avg}|\}>1\\ 1, & others\end{cases} \tag{4-25}$$

该定义表明，群体适应度方差σ^2反映的是粒子群中所有粒子的“收敛”程度。σ^2越小，则粒子群越趋于收敛；反之，粒子群则处于随机搜索阶段。

定义2　设粒子群中某个粒子在t时刻的位置为$x(t)$，p为搜索空间内的任意位置，则粒子收敛定义如下(Van Den Bergh，2002)：

$$\lim_{t\to\infty}x(t)=p \tag{4-26}$$

该定义表明，粒子的收敛是指粒子最终停留在搜索空间内某一固定位置p。

定理　如果粒子群优化算法陷入早熟收敛或者达到全局收敛，粒子群中的粒子将聚集在搜索空间的一个或几个特定位置，这时群

体适应度方差σ^2等于零。

该定理给出了粒子群优化算法群体适应度方差与收敛状态之间的关系。对于粒子群中的任意粒子，其最终收敛位置将是整个粒子群找到的全局极值。如果粒子群找到的全局极值只有一个，那么所有粒子都会聚集到该位置；如果全局极值不止一个，那么粒子将随机聚集在这几个全局极值位置。全局极值是所有粒子在算法运行过程中找到的最佳粒子位置，该位置并不一定就是搜索空间中的全局最优点。因此仅凭群体适应度方差等于零不能区别早熟收敛与全局收敛，还须进一步判断算法此时得到的最优解是否为理论全局最优解或者期望最优解f_d。如果此时已经得到全局最优，则可认为算法达到全局收敛；反之，则表明算法陷入局部最优。

根据公式(4-22)和公式(4-23)，粒子下一时刻的位置由当前位置与当前速度共同决定，速度方向决定粒子前进方向，速度大小决定移动距离。其中，粒子当前速度包括3个因素：原来的速度、个体极值p_{id}与全局极值p_{gd}。全局极值p_{gd}是算法目前找到的最优解。如果算法出现早熟收敛，全局极值p_{gd}一定是局部最优解。结合公式(4-22)，如果此时通过变异操作改变全局极值p_{gd}，就可以改变粒子的前进方向，从而让粒子进入其他区域进行搜索，在其后的搜索过程中，算法就可能发现新的个体极值p_{id}以及全局极值p_{gd}。如此循环，算法就可以找到全局最优解。

考虑到粒子在当前p_{gd}的作用下可能发现更好的位置，因此新算法将变异操作设计为一个随机算子，即对满足变异条件的p_{gd}按一定的概率p_m变异。p_m计算公式如下：

$$p_m=\begin{cases}k，\sigma^2<\sigma_d^2 and f(p_{gd})<f_d\\0，others\end{cases} \tag{4-27}$$

其中，k可以取[0.1，0.3]之间的任意值，f_d可以设置为理论最优值。这里考虑的是分类精度“最大化”情况。

对于p_{gd}的变异操作，将采用增加随机扰动的方法，设p_{gd}^k为p_{gd}的第k维取值，η是服从Gauss(0，1)分布的随机变量，则

$$p^k{}_{gd}=p^k{}_{gd}\cdot(1+0.5\cdot\eta) \tag{4-28}$$

变异操作在迭代中不断缩小种群搜索空间，当出现早熟收敛时，能使粒子跳出先前找到的最优值位置，在更大的空间中开展新的搜索，同时变异操作也保持了种群的多样性。

4.2.3　土地覆盖分类模型构建

SVM做分类预测时需要调节相关参数才能得到理想的预测分类准确率，通常采用的方法为交叉验证(CV，Cross Validation)，主要思想是让惩罚参数、核函数参数在一定的范围内取值，最终取使得训练集验证分类准确率最高的那组作为最佳参数，但有一个问题就是可能会有多组参数对应于最高的验证分类准确率，此时便出现对于同时达到最高验证分类准确率参数的取舍问题。本书粒子表示为SVM的主要参数，在此将对训练集进行CV意义下的分类精度作为AMPSO中的适应度函数，如公式(4-29)所示：

$$f=\frac{cc}{cc+uc}\times 100\% \tag{4-29}$$

这里cc和uc分别代表正确样本数和错分样本数。

利用AMPSO对SVM的参数寻优，其流程如图4-3所示。

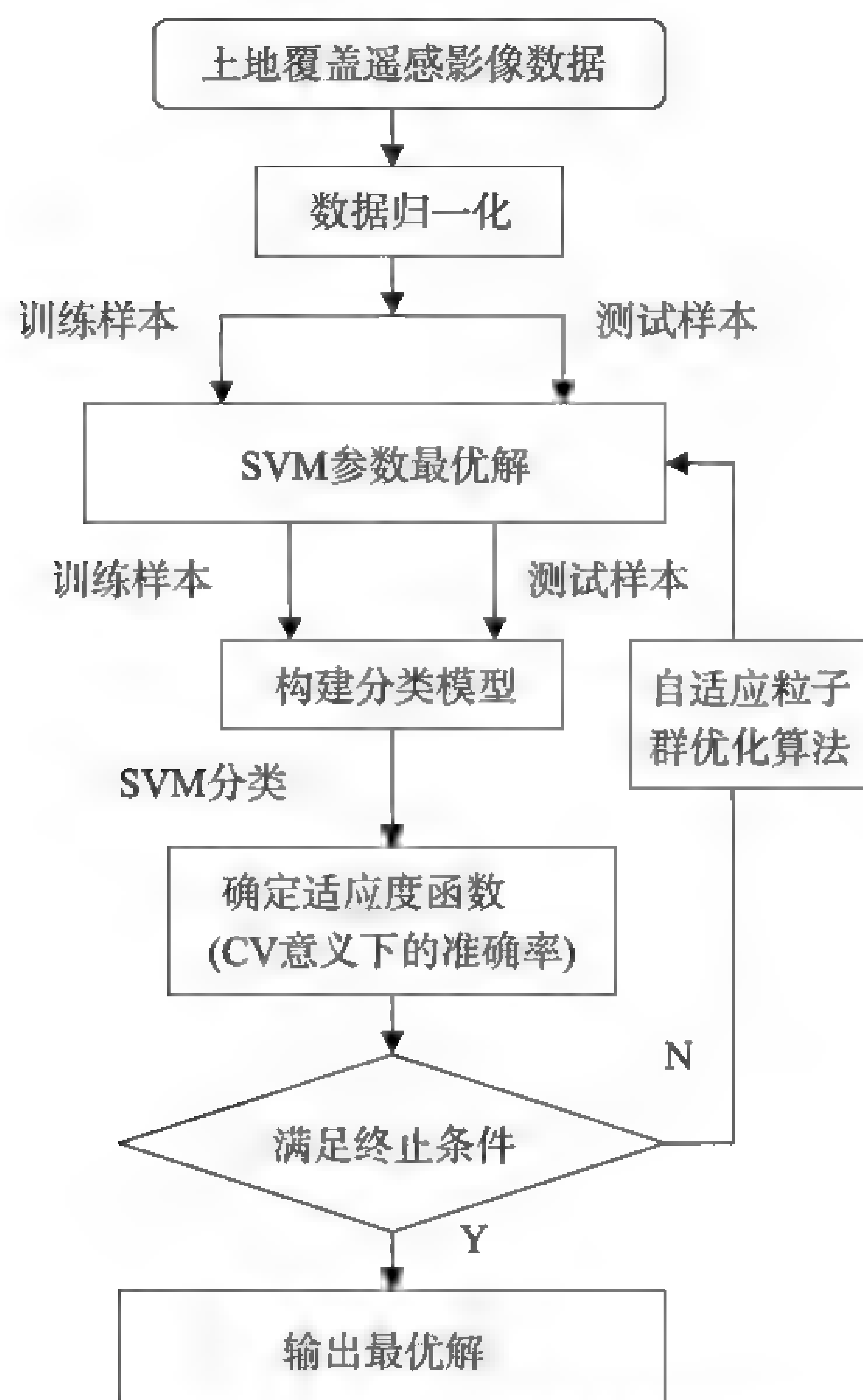


图4-3　AMPSO优化SVM参数的算法流程图

AMPSO-SVM算法的具体步骤如下。

Step1：随机初始化粒子群中粒子的位置和速度。

粒子由SVM的主要参数构成。SVM4种核函数中RBF核具有较宽的收敛性，不受维数以及样本数量的严格限制，本模型选择RBF作为分类依据函数。由此初始化粒子包括惩罚参数c和RBF核函数参数γ两部分。

Step2：将粒子p_b设置为当前位置，p_g设置为初始群体中最佳粒子位置。

Step3：判断算法是否满足收敛条件，如果满足则执行Step9，否则执行Step4。

粒子群优化算法最终达到收敛位置时，整个粒子群将获得全局极值，利用群体适应度方差σ^2、全局极值$f(p_g)$与理论极值f_d比较作为全局收敛判定准则；将SVM得到的分类准确度作为粒子的适应度函数，计作f，如公式(4-29)所示；收敛条件为同时满足公式(4-30)和公式(4-31)：

$$\sigma^2=\sum_{i=1}^{n}\left[\frac{f_i-f_{avg}}{f}\right]^2=0 \tag{4-30}$$

$$f(p_g)\geqslant f_d \tag{4-31}$$

其中，f_i为第i个粒子的适应度，f_{avg}为粒子群目前的平均适应度。

Step4：速度更新，位置更新。根据公式(4-22)和公式(4-23)更新速度和位置。

Step5：据公式(4-24)和公式(4-25)计算群体适应度方差σ^2，并计算$f(p_g)$。

Step6：根据公式(4-27)计算变异概率p_m。

Step7：产生随机数$r\in[0,1]$，如果$r<p_m$，按公式(4-28)执行变异操作，否则，转向Step8。

Step8：判断Step3收敛准则是否满足，如果满足，执行Step9，否则执行Step4。

Step9：输出p_g，算法结束。

4.3 实验结果与分析

4.3.1 实验影像选取

为了验证AMPSO-SVM模型的有效性，本文选择2006年5月17日获取行列号115-30多光谱Landsat-5 TM遥感影像。由于幅面有限，图形比例尺较小，难以从视觉上直观反映各种分类算法的性能。为此，从研究区原始TM遥感影像图中切割出典型区域，从微观上比较各类算法的分类结果。图4-4为研究区T543假彩色合成。实验为了保持研究区各种地物的原始光谱特征，提高自动分类的准确性和科学性，影像未作数字增强处理。此外，由于实验重点在于土地覆盖遥感影像信息提取算法研究，所有影像未作投影变换，仍采用默认的UTM投影。

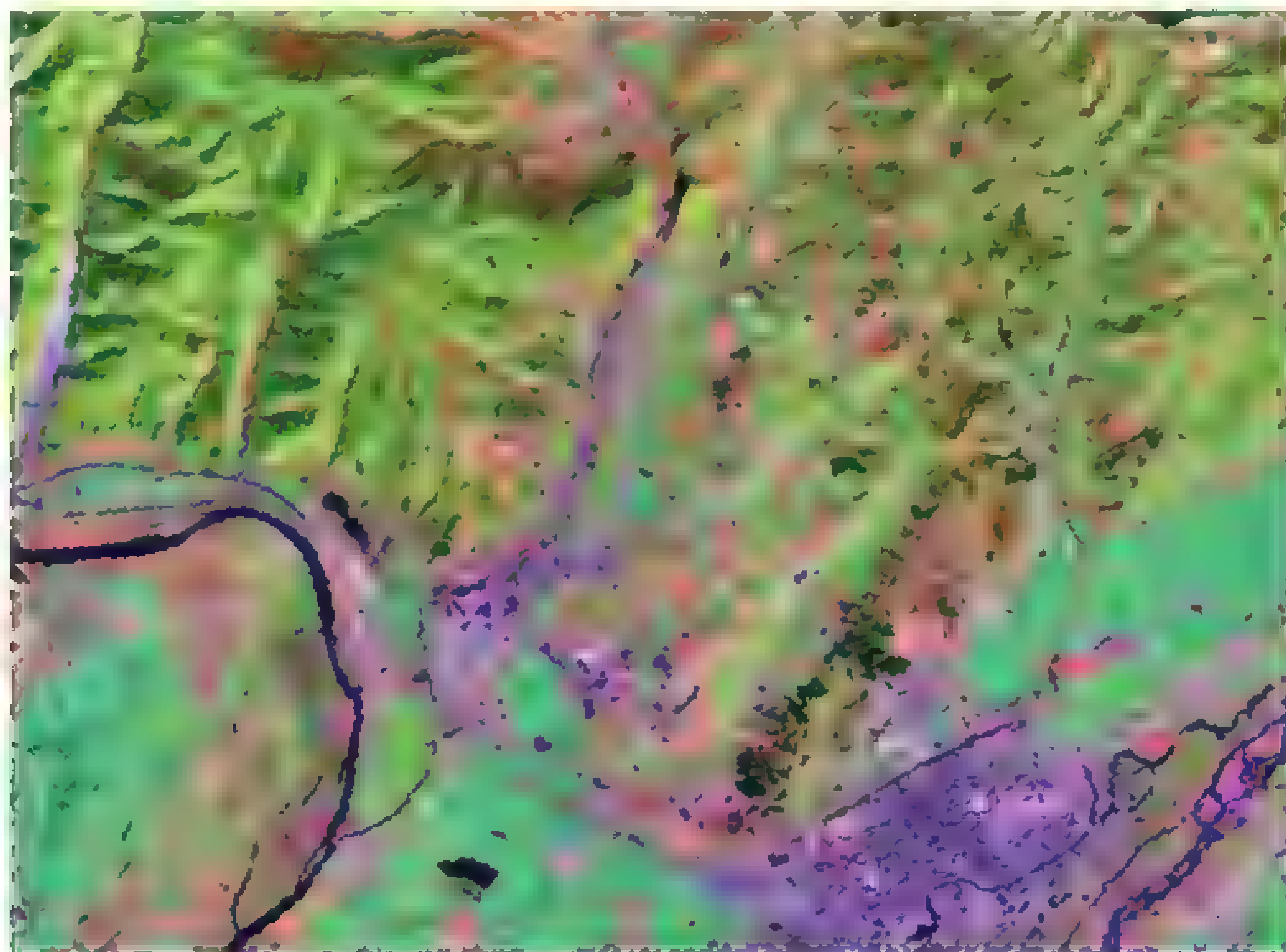


图4-4　研究区原始影像(5、4、3波段合成)

4.3.2　特征选取及样本集表示

特征提取与特征选择是模式识别中的关键技术之一。如第3章介绍，本文提取了8个特征，包括TM图像的6个波段(1～5波段，7波段)、PCA变换的第一主分量、植被指数(NDVI)。

据第3章介绍研究区土地覆盖分类系统，本章采用一级分类，研究区共分为5大类，即建筑用地、农田、水体、其他和森林，分别用类别代号ω_1～ω_5表示，从试验区遥感影像上应用简单随机采样方法选取共计2 059个训练样本和1 779个验证样本(详见表4-1)，用于训练分类算法和评价分类精度。让每个样本集、测试集独立于训练集。整个实验在Intel酷睿3.0 GHz CPU、内存为2GB的台式计算机上进行，算法采用Matlab 9.0及ENVI4.5软件编程实现，SVM采用LIBSVM工具。

表4-1　类别及样本数量

类别代号	类别名称	训练样本	验证样本
ω_1	建筑用地	335	418
ω_2	农田	550	371
ω_3	水体	125	237
ω_4	其他	367	389
ω_5	森林	682	364
类别及样本总数	5	2 059	1 779

4.3.3　核函数的选取

除了公式(4-17)和公式(4-21)介绍的常用核函数外，还包括指数径向基核函数、小波核函数、傅里叶核函数等其他一些核函数，但应用相对较少。在训练SVM的过程中，多项式核函数耗费时间更多，而且它比径向基核函数有更多的参数需要调整和设置，当多项

式的次数很高时，它可能趋近于零或无穷[19]；Sigmoid核函数对某些参数的表现与径向基函数相似，但它对某些参数是无效的[20]。许多前人的研究成果表明径向基核函数比多项式核函数具有更好的分类性能，特别是当样本数据量很大时，径向基核函数的收敛性能强于其他核函数[21]~[23]。

4.3.4 实验参数及精度评价指标

1. 主要参数

除了AMPSO寻优获得的SVM惩罚参数与核函数参数外，分类模型还包括SVM的初始化参数值设为3；种群规模为30；最大迭代次数为100；学习因子c_1设为1.6；学习因子c_2设为1.5。

2. 精度评价指标

分类精度是评价分类器性能的重要指标，目前的分类精度评价方法很多，包括Mapping Accuracy Index[24]，Mean Accuracy Index[25]，Classification Success Index[26]等。误差矩阵是一个常用的遥感影像分类精度描述模型，本文采用基于误差矩阵的总体分类精度、用户精度、生产精度及Kappa系数的分类精度评价方法来评价AMPSO-SVM算法的分类性能，分别表示如下[27]~[28]，其中p_{ij}表示分类数据类型中第i类和实测数据类型中第j类所占的组成部分；$p_{i+}=\sum_{i=1}^{n}p_{ij}$为分类所得到的第$i$类的总和；$p_{+j}=\sum_{i=1}^{n}p_{ij}$为实际观测的第$j$类的总和；$p$为样本总数。

(1) 总体分类精度

$$p_{+j}=\sum_{i=1}^{n}p_{ij} \tag{4-32}$$

P_0是具有概率意义的一个统计量，表述的是对每一个随机样本所分类的结果与地面对应区域的实际类型相一致的概率。

(2) 用户精度(对于第i类)

$$p_{ui} = p_{ii}/p_{i+} \tag{4-33}$$

表示从分类结果中任取一个随机样本，其所具有的类型与地面实际类型相同的条件概率。

(3) 生产精度(对于第j类)

$$p_{Aj} = p_{jj}/p_{+j} \tag{4-34}$$

表示相对于地面获得的实际资料中的任意一个随机样本，分类图上同一地点的分类结果与其相一致的条件概率。

(4) Kappa系数

总体精度、用户精度和制图精度从不同的侧面反映了分类精度的统计估计，Kappa系数分析采用了另一种离散的多元技术，是一种用来测定两幅图之间吻合度或精度的指标。其公式表达如下[27]：

$$Kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}x_{+i})} \tag{4-35}$$

4.3.5　实验结果与比较

第一个试验是比较SVM的Sigmoid核函数和径向基核(Radial Basis Function)分类性能。以SVM-Sigmoid和SVM-RBF分类器，在固定参数c 100，γ 0.143条件下对待分类遥感数字集(如表4-2)进行分类，并与传统最大似然法(Maximum Likelihood)在生产精度、用户精度、总精度，以及Kappa系数评价指标下进行比较，

其结果如表4-2所示。首先，使用SVM方法分类的精度明显高于Maximum Likelihood方法，其中SVM-Sigmoid方法得到的总分类精度比Maximum Likelihood高出0.68%，而SVM-RBF方法得到的分类精度比Maximum Likelihood高出3.15%，SVM-RBF方法得到的Kappa系数分别比SVM-Sigmoid方法和Maximum Likelihood高出0.031 4和0.039 8。因此，可以得出SVM-RBF方法的分类性能最优，本文采用RBF核函数。

第二个实验采用传统PSO和AMPSO优化算法对SVM的c和γ寻优。图4-5和图4-6分别为两种优化算法获得的适应度曲线(以分类精度为适应度函数)。从结果中可以看出，带变异算子的粒子群优化算法能够及时跳出局部极小值，收敛效果更优。同时表4-3列出两个分类模型所获得的最优c和γ参数值，以及总精度和Kappa系数，其中由PSO-SVM分类模型所得到c和γ的值为166.942 3和1.366；AMPSO-SVM分类模型所得到c和γ的值246.789 1和0.134，总精度从PSO-SVM的91.50%提高到AMPSO-SVM的93.59%，Kappa系数由0.890 3提高为0.917 5，结果明显优于SVM的手工设置值100和0.143所得到的结果，这些均表明AMPSO-SVM模型能够有效提高遥感影像的分类精度。

表4-2 不同分类器分类精度比较

分类方法	5种地物类型生产精度(%)					5种地物类型用户精度(%)					总精度(%)	Kappa系数
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_1	ω_2	ω_3	ω_4	ω_5		
SVM-RBF	98.33	100.00	89.03	56.30	92.58	98 80	94.16	99.53	92.80	64.68	87.07	0.837 2
SVM-Sigmoid	99.28	99.73	83.54	53.73	85.99	93.68	95.85	100.00	81.64	63.10	84.60	0.805 8
Maximum Likelihood	98.56	99.73	88.61	63.24	70.05	98.56	78.72	99.53	92.48	61.69	83.92	0.797 4

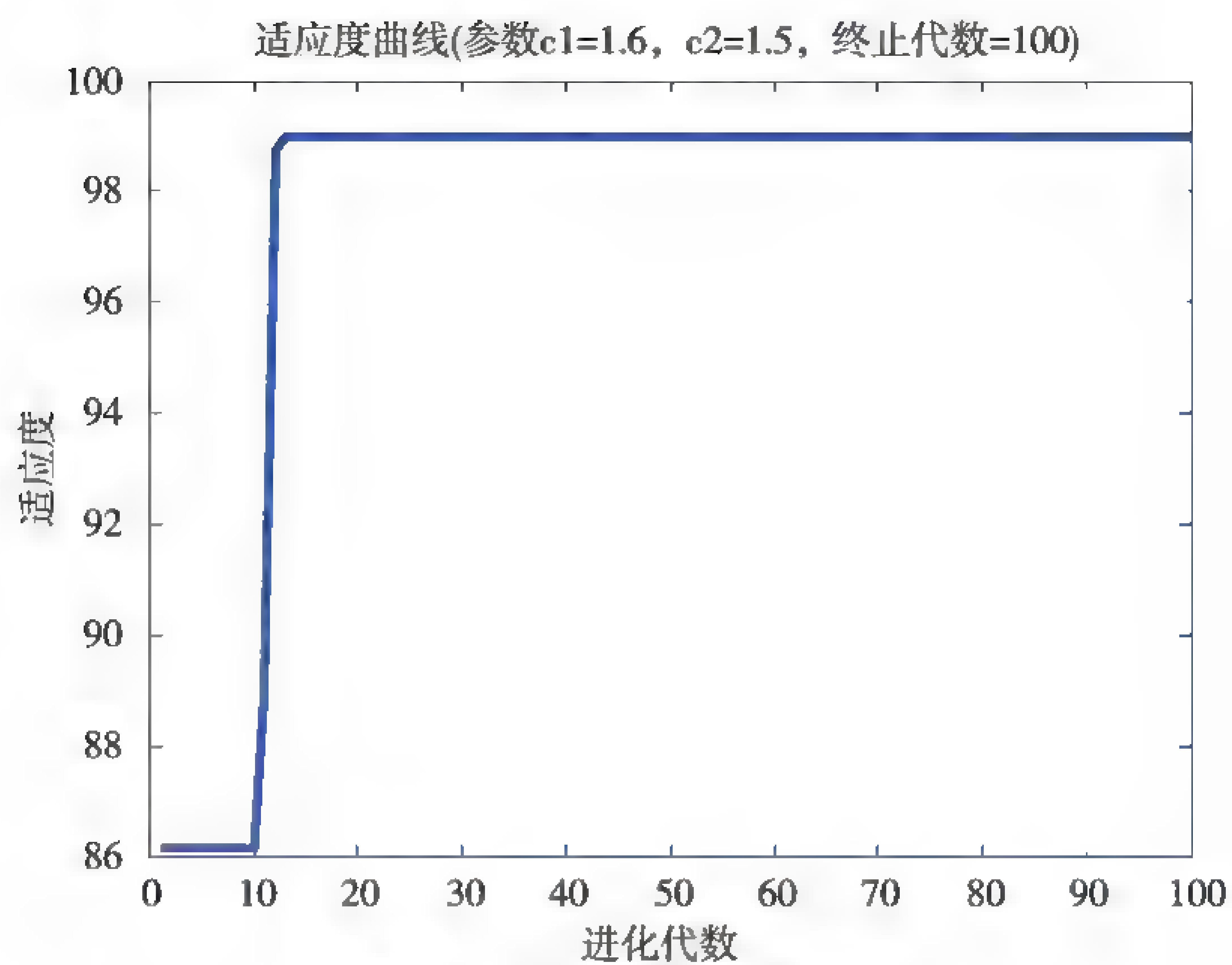


图4-5　PSO-SVM适应度曲线

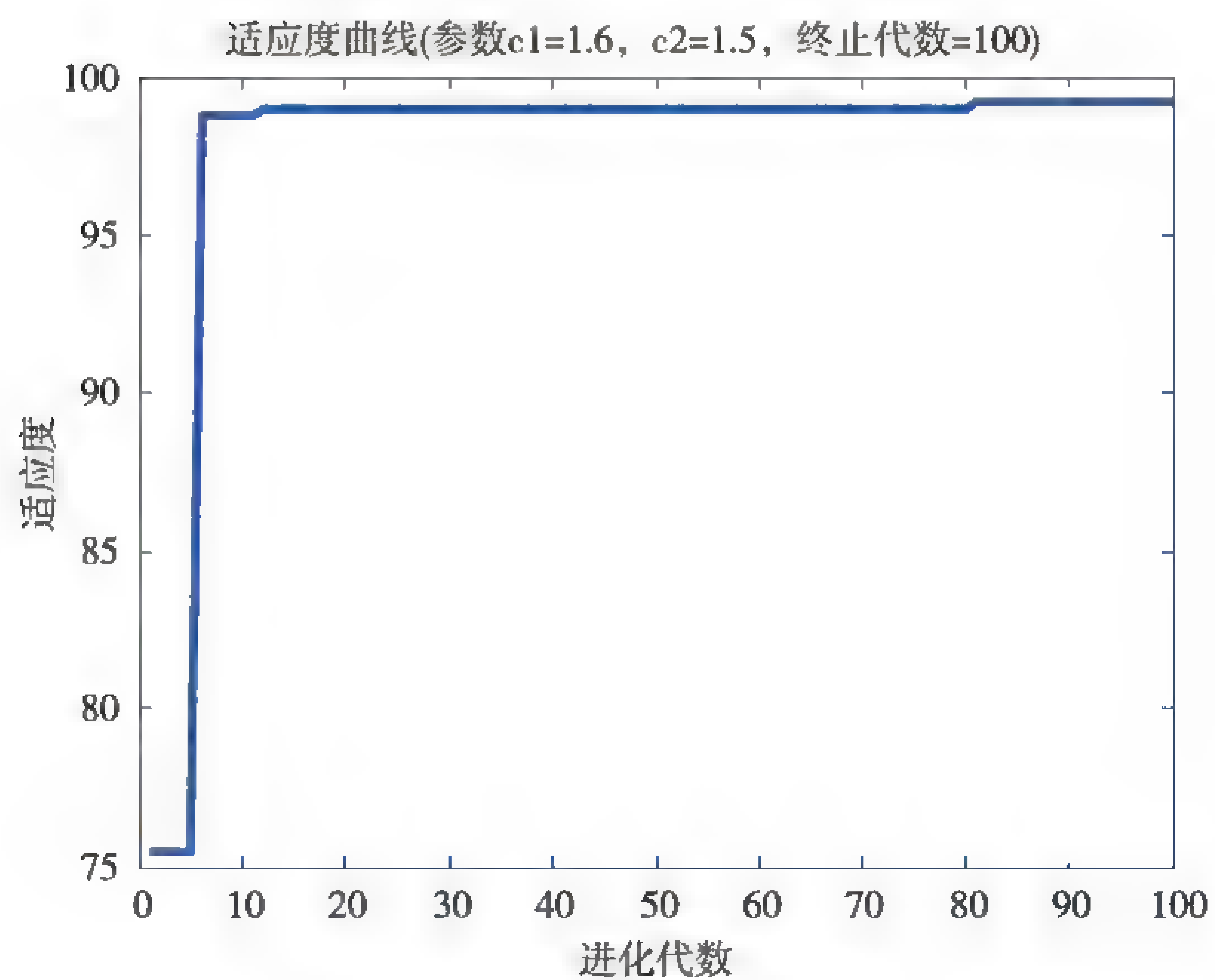
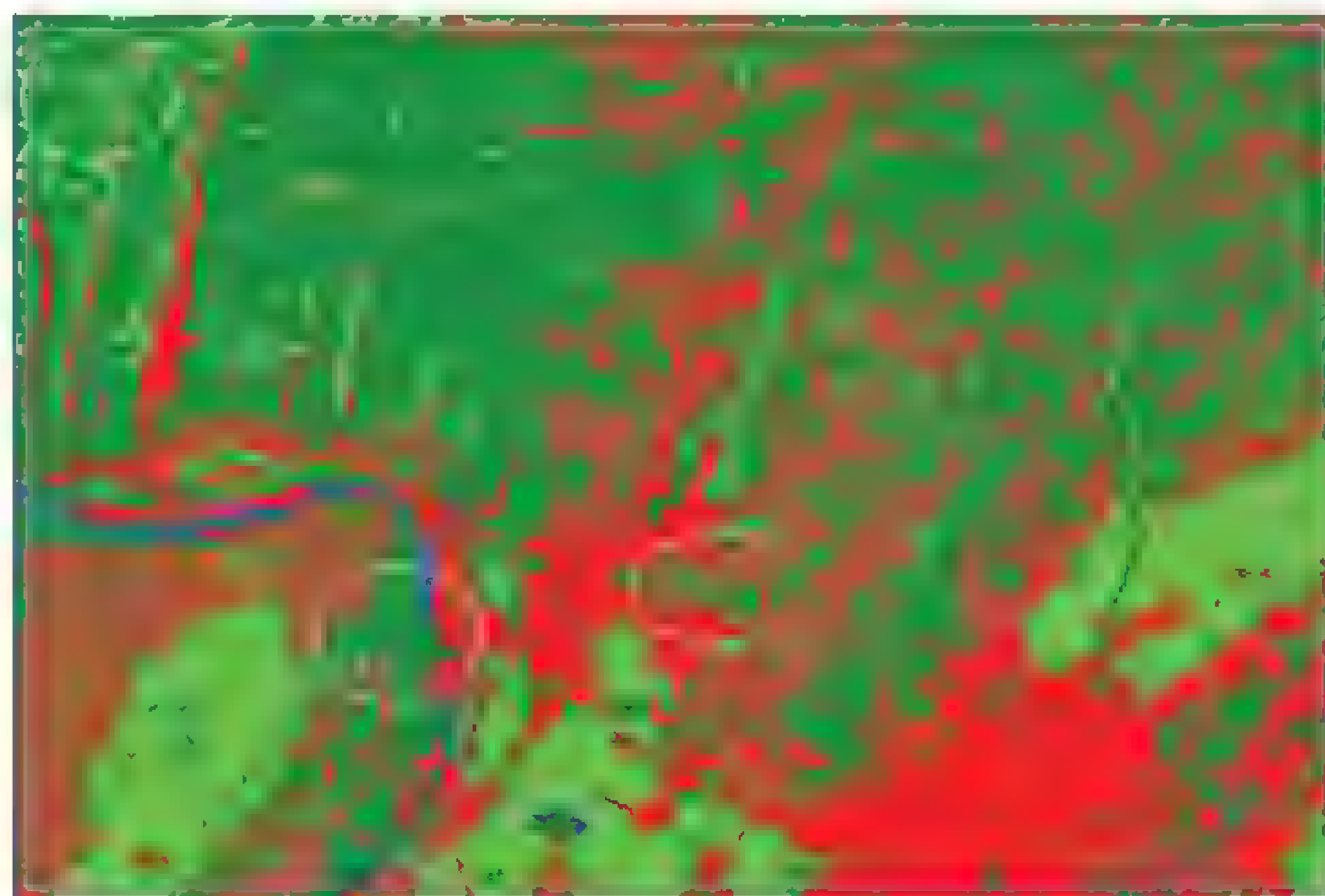


图4-6　AMPSO-SVM适应度曲线

试验最后将所提出AMPSO-SVM模型应用于研究区土地覆盖分类试验。为了比较其分类性能，分别与Maximum Likelihood法、传统SVM分类法、PSO-SVM分类法对比。分类结果如图4-7所示，其中图4-7(a)为最大似然法分类结果，分类后获取的地类斑块破碎化程度大，零碎地类斑块多，存在严重的森林与植被的混分现象；图4-7(b)采用传统SVM分类方法对影像的分类结果，图4-7(c)和(d)分别采用PSO-SVM和AMPSO-SVM方法的分类结果，从图中可以直观的看出图4-7(d)的分类效果明显好于其他3种方法，尤其是针对森林和耕地的区分。

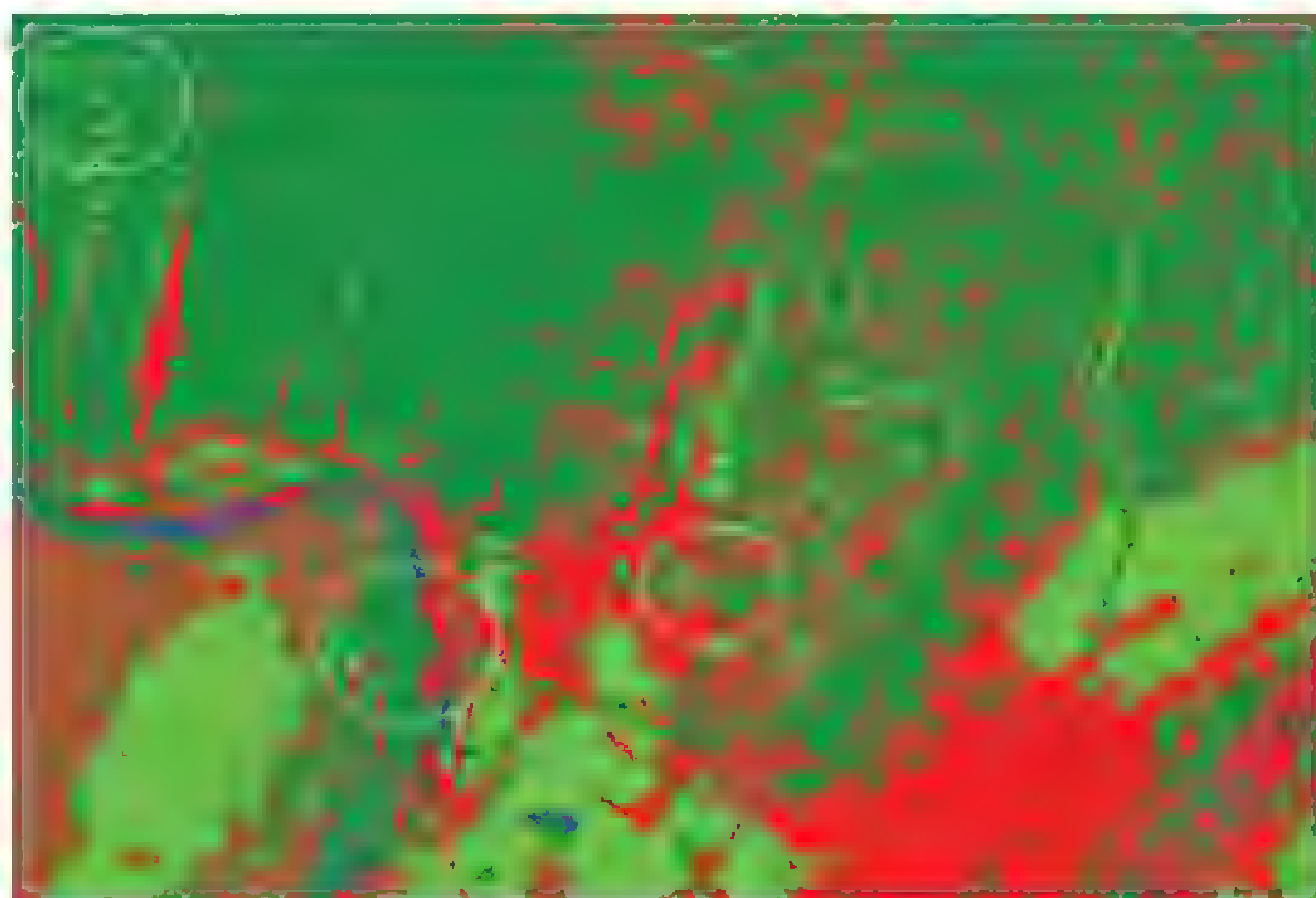
表4-3　三种分类方法参数及分类精度、Kappa系数比较

分类模型	惩罚参数c	核函数参数γ	分类精度(%)	Kappa系数
SVM	100	0.143	87.07	0.837 2
PSO-SVM	166.942 3	1.366	91.50	0.890 3
AMPSO-SVM	246.789 1	0.134	93.59	0.917 5

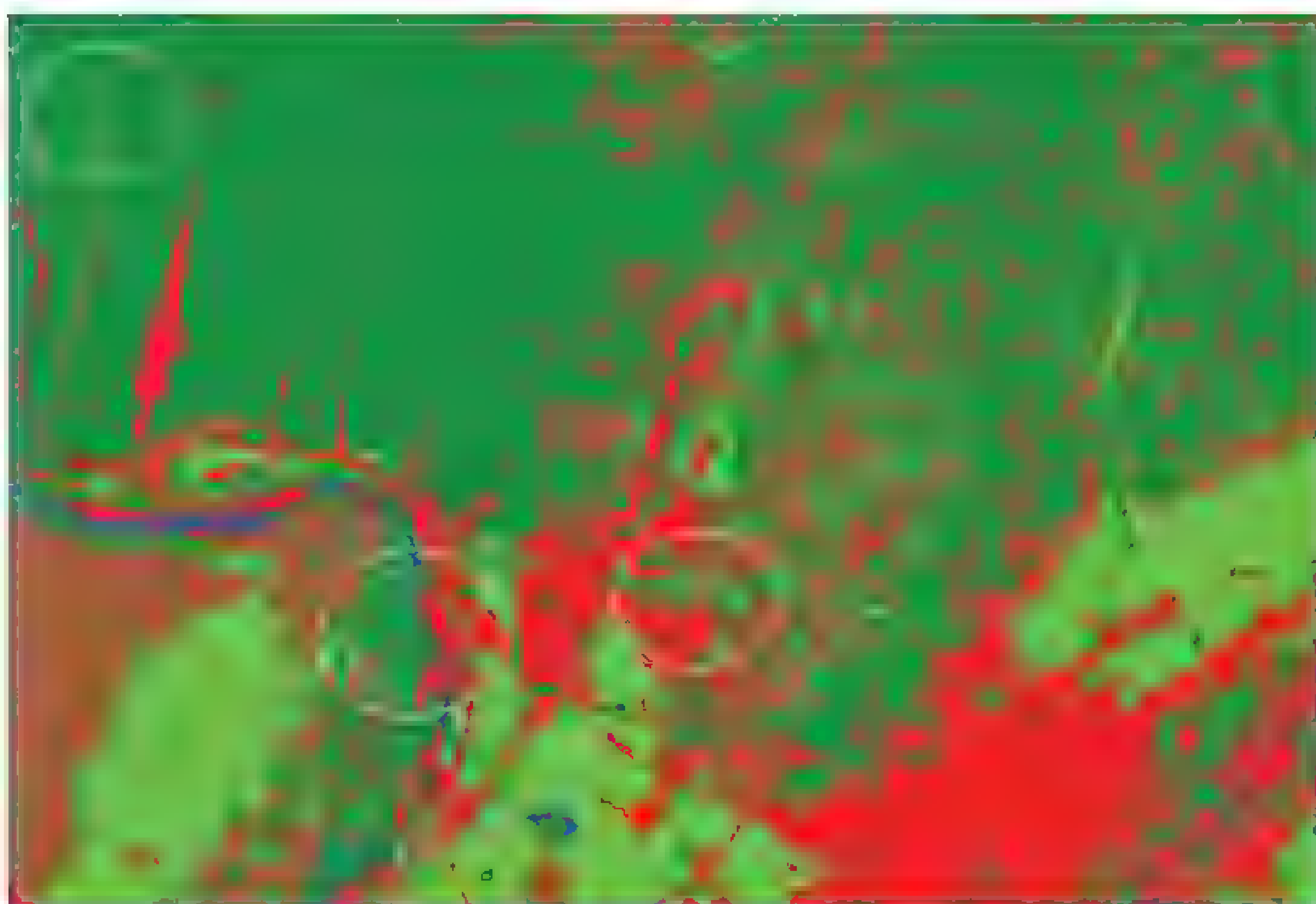


(a) Maximum Likelihood 分类结果

图4-7　四种分类方法的分类结果

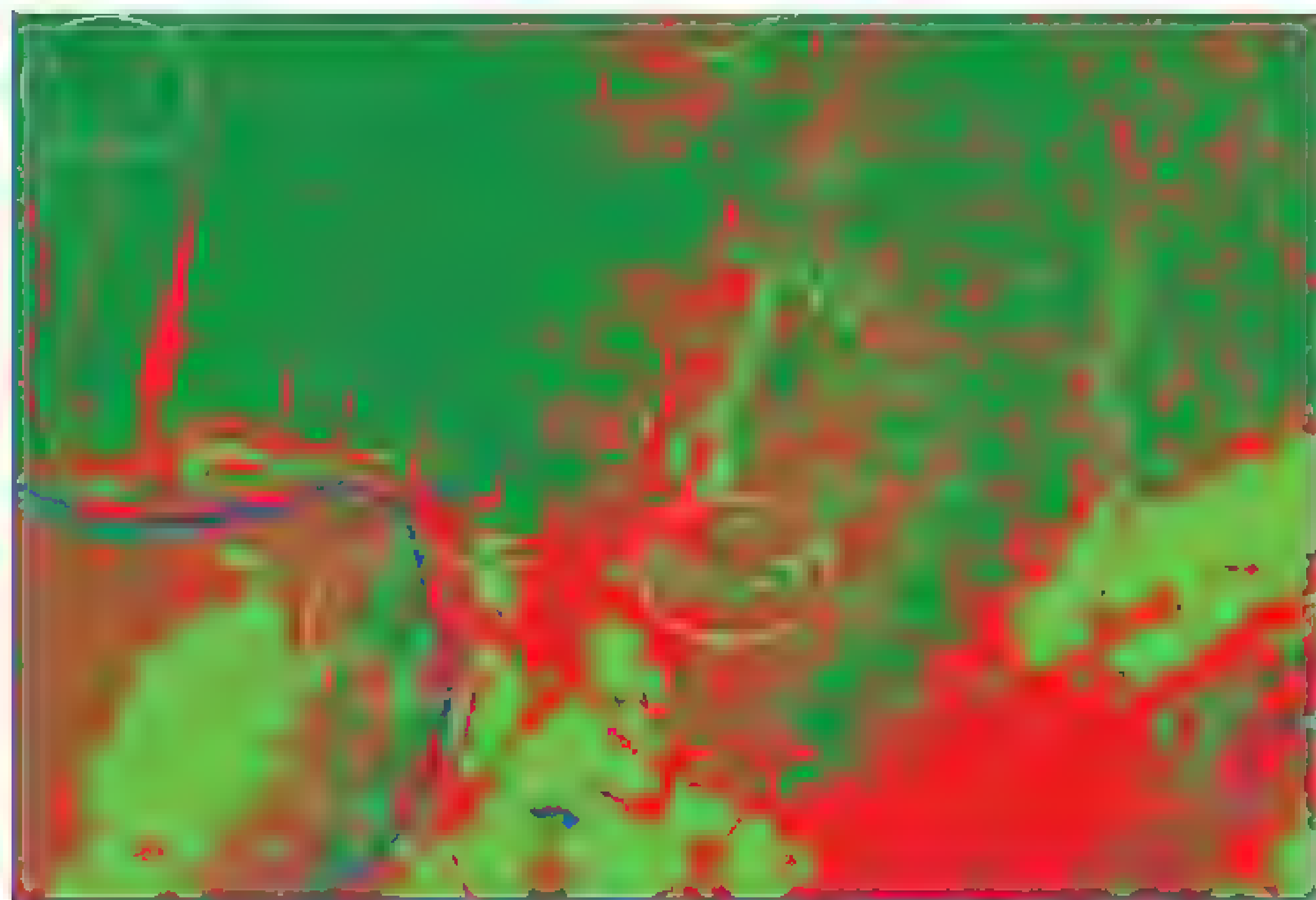


(b) SVM分类结果



(c) PSO-SVM分类结果

图4-7　四种分类方法的分类结果(续)



(d) AMPSO-SVM分类结果



图4-7　四种分类方法的分类结果(续)

4.4　本章小结

本文针对传统PSO优化SVM参数存在早熟收敛、后期迭代效率不高从而造成参数寻优不准确的缺点，提出了一种基于自适应变异粒子群算法的SVM参数优化模型(AMPSO-SVM)。自适应粒子群优化算法是以粒子群的群体适应度方差σ^2和全局极值与理论最优值的比较作为粒子群优化算法是否陷入局部极值的评价指标，引入变异算子使得算法能够及时跳出局部极值进而获得全局最优解。AMPSO-SVM能够克服传统SVM分类模型参数选择的主观性，同时可以快速摆脱局部搜索的束缚，实验表明，该模型能有效提高遥感影像分类的精度。

参考文献

[1] Vapnik V.N.. *The Nature of Statistical Learning Theory*[M]. New York: Springer-Verlag, 1995.

[2] Warner T.A., Nerry F.. *Does Single Broadband or Multispectral Thermal Data Add Information for Classification of Visible, Near- and Shortwave Infrared Imagery of Urban Areas*[J]. International Journal of Remote Sensing, 2009, 30(9): 2155-2171.

[3] Lardeux C., Frison P.L., Tison C., Souyris J.C., Stoll B., Fruneau B., Rudant J.P.. *Support Vector Machine for Multifrequency SAR Polarimetric Data Classification*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(12): 4143-4152.

[4] Knorn J., Rabe A., Radeloff V.C., Kuemmerle T., Kozak J., Hostert P.. *Land Cover Mapping of Large Areas Using Chain Classification of Neighboring Landsat Satellite Images*[J]. Remote Sensing of Environment, 2009, 113(5): 957-964.

[5] Heikkinen V., Tokola T., Parkkinen J., Korpela I., Jaaskelainen T.. *Simulated Multispectral Imagery for Tree Species Classification Using Support Vector Machines*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(3): 1355-1364.

[6] Kavzoglu T., Colkesen I.. *A Kernel Functions Analysis for Support Vector Machines for Land Cover Classification*[J]. International Journal of Applied Earth Observation and Geoinformation, 2009, 11: 352-359.

[7] 王睿. 关于支持向量机参数选择方法分析[J]. 重庆师范大学学报(自然科学版), 2007, 24(2): 36-38.

[8] 吴渝, 向浩宇, 刘群. 一种基于网格的最近邻SVM新算法[J]. 重庆邮电大学学报(自然科学版), 2008, 20(6): 706-709.

[9] 李京华, 张聪颖, 倪宁. 基于参数优化的支持向量机战场多目标声识别[J]. 探测与控制学报, 2010, 32(1): 1-5.

[10] Hsu C.W., Lin C.J.. *A Simple Decomposition Method for Support Vector Machine*[J]. Machine Learning, 2002, 46(3): 219-314.

[11] LaValle S.M., Branicky M.S.. *On the Relationship between Classical Grid Search and Probabilistic Roadmaps*[J]. International Journal of Robotics Research, 2002, 23(8): 673-692.

[12] Fröhlich H., Chapelle O.. *Feature Selection for Support Vector Machines by Means of Genetic Algorithms*[C]. The 15th IEEE International Conference on Tools with Artificial Intelligence. USA: Sacramento, CA, 2003: 142-148.

[13] Zheng C.H., Jiao L.C.. *Automatic Parameters Selection for SVM Based on GA*[M]. The 5th World Congress on Intelligent Control and Automation. Piscataway, NJ: IEEE Press, 2004: 1869-1872.

[14] Vahid R., Ata E., Reza G.. *Application of the PSO-SVM Model for Recognition of Control Chart Patterns*[J]. ISA Transactions, 2010, 49: 577-586.

[15] Huang C.L., Dun J. F.. *A Distributed PSO-SVM Hybrid System with Feature Selection and Parameter Optimization*[J]. Applied Soft Computing, 2008, 8: 1381-1931.

[16] 丁胜, 袁修孝, 陈黎. 粒子群优化算法用于高光谱遥感影像分类的自动波段选择[J]. 测绘学报, 2010, 39(3): 257-263.

[17] 陈仕涛, 陈国龙, 郭文忠. 基于粒子群优化和邻域约简的入侵检测日志数据特征选择[J]. 计算机研究与发展, 2010, 47(7): 1261-1267.

[18] 吕振肃, 侯志荣. 自适应变异的粒子群优化算法[J]. 电子学报, 2004, 32(3): 416-420.

[19] Jae H.M., Young C.L.. *Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters*[J]. Expert Systems with Applications, 2005, 28: 603-614.

[20] Vapnik V.N.. *Statistical Learning Theory*[M]. New York: Wiley, 1998.

[21] Kim K.J.. *Financial Time Series Forecasting Using Support Vector Machines*[J]. Neurocomputing, 2003, 55: 307-319.

[22] Huang Z., Chen H., Hsu C.J., Chen W.H.. *Credit Rating Analysis with Support Vector Machine and Neural Networks: A Market Comparative Study*[J]. Decision Support Systems, 2004, 37: 543-558.

[23] 梅建新, 段汕, 秦前清. 基于支持向量机的特定目标检测方法[J]. 武汉大学学报(信息科学版), 2004, 29(10): 912-915.

[24] Short N.M.. *The Landsat Tutorial Workbook-Basics of Satellite Remote Sensing*[M]. Greenbelt, Md., Goddard Space Flight Center, NASA Reference Publication, 1982: 1078.

[25] Hellden U.. *A Test of Landsat-2 Imagery and Digital Data for Thematic Mapping Illustrated by an Environmental Study in Northern Kenya*[C]. Sweden: Lund University Natural Geography Institute Report,

1980: 47.

[26] Koukoulas S., Blackburn G.A.. *Introducing New Indices for Accuracy Evaluation of Classified Images Representing Semi-natural Woodland Environments*[J]. Photogrammetric Engineering and Remote Sensing, 2001, 67: 499-510.

[27] 赵英时. 遥感应用分析原理与方法[M]. 北京:科学出版社, 2003.

[28] 李建平. 基于FNMPSVM模型和图像分割的盐碱地信息提取研究[C]. 中国科学院, 2007.

第5章

基于模糊聚类的半监督支持向量机土地覆盖分类方法研究

5.1 概述

遥感信息是地球表面地物的综合反映，地物种类繁多且地物之间存在界限的模糊性和重叠性[1]，由此造成遥感信息具有很强的不确定性。模糊集理论是处理这种不确定数据的有力工具[2]~[3]，许多基于模糊统计学的算法被应用于遥感图像处理中[4]~[6]。此外，在遥感影像分类过程中人为选择样本少且代表性不好，也是影响分类精度的又一关键问题。

半监督学习是在较少的训练标签样本的基础上，同时利用大量、廉价的未标记样本，通过挖掘未标记样本中所蕴涵的各待分类类型在特征空间中的固有结构信息，来对已标记样本可能因代表性不好而造成的拟合分类器有偏差情况进行矫正[7]~[8]。如果能把大量无标签样本所包含的数据特征加入到学习算法的设计中，就可以弥补单个监督分类器的不足，获得更好的分类效果。

本章一方面利用自训练半监督学习算法与第4章介绍优化的SVM结合构建半监督分类模型；另一方面在自训练算法的未标记样本标注过程中，利用模糊聚类理论，比较3种可靠无标签样本标注方法，主要目的在于控制错误类别样本的标注，进而提高分类器的分类精度。

5.2 自训练半监督学习

5.2.1 无标签样本的重要性

半监督学习的基本设置是给定一个来自某一未知分布的有标记

样例集$L=\{(x_1, y_1), (x_2, y_2), \cdots, (x_{|L|}, y_{|L|})\}$，以及一个无标记样例集$U=\{x'_1, x'_2, \cdots, x'_U\}$，期望学习一个函数$f: X \to Y$可以准确地对样例$x$预测其标记$y$[9]。

Miller和Uyar[10]从数据分布估计的角度给出了一个直观的分析，假设所有数据服从某个由L个高斯分布混合而成的分布：

$$f(x|\theta)=\sum_{l=1}^{L} a_l f(x|\theta) \tag{5-1}$$

其中，$\sum_{l=1}^{L} a_l = 1$为混合系数，$\theta=\{\theta_l\}$为参数。标记可视为一个由选定的混合成分m_i和特征向量x_i以概率$P(c_i|x_i, m_i)$决定的随机变量。于是，根据最大后验概率假设，最优分类由以下公式给出：

$$h(x)=\arg\max_k \sum_j P(c_i=k|m_i=j, x_i)P(m_i=j|x_i) \tag{5-2}$$

这样，学习目标就变成了利用训练样例来估计$P(c_i=k|m_j=j, x_i)$和$P(m_i=j|x_i)$。这两项中的第一项与类别标记有关，而第二项并不依赖于样例的标记，如果有大量的无标记样例可用，则意味着能够用于估计第二项的样例数显著增多，这会使得第二项的估计变得更加准确，从而导致公式(5-2)更加准确。也就是说，分类器的泛化能力得以提高。因此，无标记样例的价值就在于它们能够帮助更好地估计模型参数，从而提高模型性能。

5.2.2　自训练半监督算法

自训练方法是半监督学习比较常用的方法，有时也称为self-teaching，被广泛应用于物体监测[11]、自然语言处理[12]~[13]，以及遥感影像分类[14]。自训练方法首先用有标签样本训练一个分类器，然后用此分类器对所有无标签样本进行分类，并给每个无标签样本标

上类别标签和相应的置信度；再将置信度高的样本连同它的类别标签合并到训练集中继续训练分类器；重复上述过程直至结束条件满足。

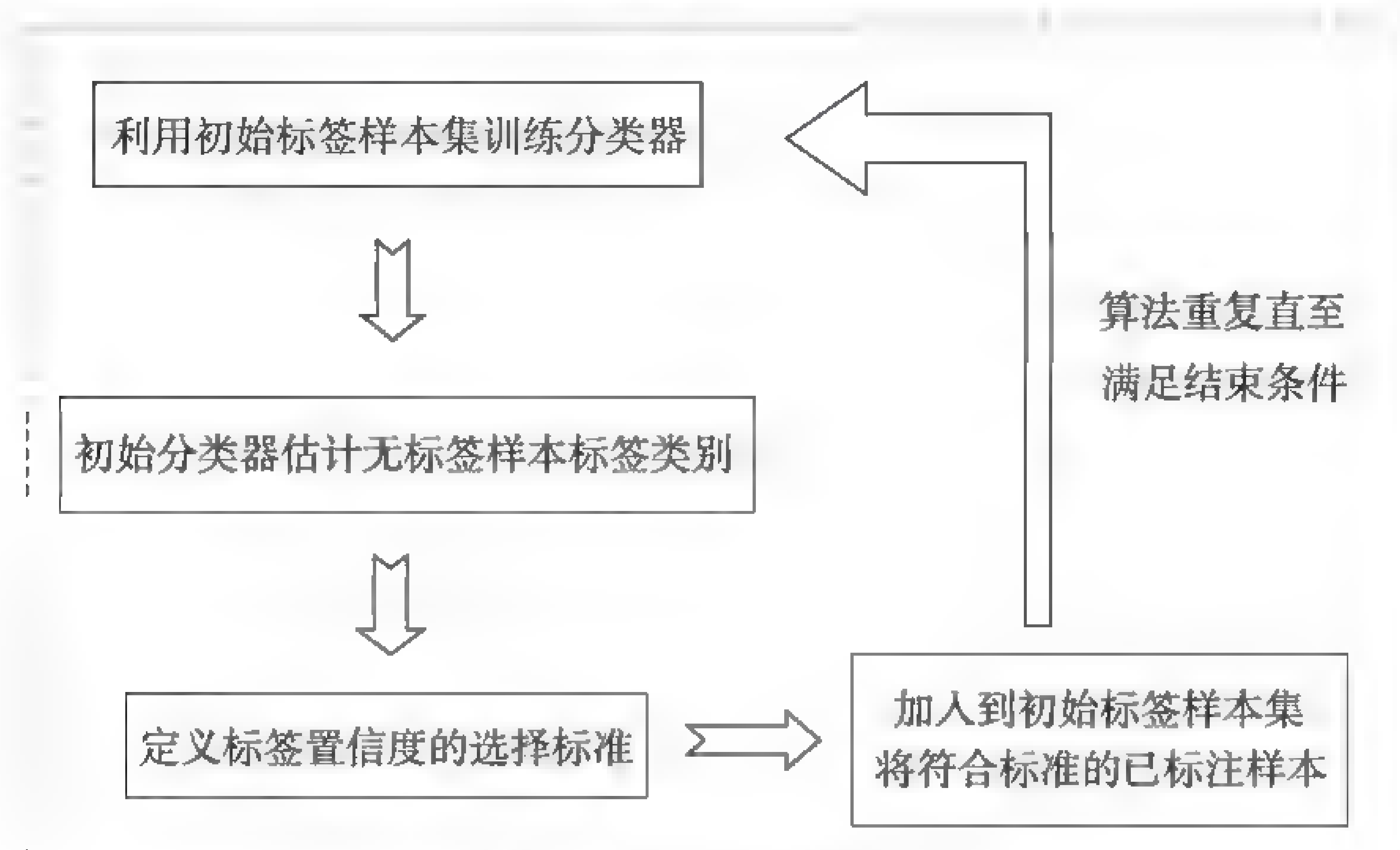


图5-1 自训练半监督算法

设已标注样本集合L，未标注样本集合为M，类别数为C，自训练算法流程如下。

Step1：从未标注样本集M中随机选出N个未标注样本组成未标注样本池M'。

Step2：迭代。

(1) 利用已标注样本集合L训练得到初始分类器；

(2) 利用初始分类器对未标注样本池M'进行标注；

(3) 从标注样本池M'中选出标注置信度大于阈值的样本加入到标注样本训练集L；

(4) 从未标注样本集M中随机选出N个未标注样本组成新的未标注样本池M'。

Step3：退出。

M是否为空或者是否满足其他退出条件，若不为空且不满足其他退出条件，返回Step2(1)，否则退出。

从Self-training算法不难看出：在初始阶段，只利用少量标注样本进行训练学习得到初始分类器，然后利用得到的初始分类器对大量未标注样本进行标注，也就是分类器是用自己的预测自我学习，那么如果某个分类器训练数据集中加入一个错误的分类信息时，所得到的分类结果也可能由于预测错误而不断加强[15]，进而扩大分类器的错误分类，这就是所谓的“错误累积”现象。

5.3　模糊聚类理论

聚类假设(Cluster Assumption)是半监督学习中的基本假设。它指出处在相同簇(cluster)中的样例有较大的可能属于同一类。

5.3.1　聚类的概念

Jain[16]在1988年关于聚类所下的定义：一个类簇内的实体是相似的，不同类簇的实体是不相似的；一个类簇是测试空间中点的会聚，同一类簇的任意两个点间的距离小于不同类簇的任意两个点间的距离；类簇可以描述为一个包含密度相对较高点集的多维空间中的连通区域，它们借助包含密度相对较低点集的区域与其他区域(类簇)相分离。事实上，聚类是一个无监督的分类，它没有任何先验知识可用。

聚类的形式描述如下[17]：

令$U=\{p_1, p_2, \cdots, p_n\}$表示一个模式(实体)集合；$p_i$表示第$i$个模式$i=\{1, 2, \cdots, n\}$；$C_t \subseteq U$；$t=1, 2, \cdots, k$；$C_t=\{p_{t_1}, p_{t_2}, \cdots, p_{t_w}\}$；$proximity(p_{ms}, p_{tr})$，其中第一个下标表示模式所属的类，第二个下标表示某类中某一模式，函数$proximity$用来刻画模式的相似性距离。若类C_t为聚类的结果，则C_t满足如下条件：

$$\cup_{t=1}^{k} C_t = U \tag{5-3}$$

对于$\forall C_m$，$C_r \subseteq U$，$C_m \neq C_r$，有：

$$C_m \cap C_r = \Phi \text{(限于刚性聚类)}$$

$$\mathrm{MIN}_{\forall p_{mu} \in C_m, \ \forall p_{rv} \in C_r, \ \forall C_m, C_r \subseteq U \& C_m \neq C_r}(proximity(p_{mu}, p_{rv})) >$$

$$\mathrm{MAX}_{\forall P_{mx}, \ P_{my} \in C_m, \ \forall C_m \subseteq U}(proximity(p_{mx}, p_{my})) \tag{5-4}$$

典型的聚类过程主要包括数据(或称样本或模式)准备、特征选择和特征提取、接近度计算、聚类(或分组)、对聚类结果进行有效性评估等步骤[18]~[20]。

5.3.2 常用聚类算法

1. K-均值聚类

K-均值(K-means)是一种非监督学习算法，由MacQueen[21]提出，用于将给定的样本集分成指定数目的聚类。K-means的聚类准则是使每一类中多模式点到该类别中心距离的平方和最小。

$$J = \sum_{i=1}^{K} J_i = \sum_{i=1}^{K}\sum_{j=1}^{N} w_{ji}\left\| X_j - C_i \right\|^2 \tag{5-5}$$

其中，J_i为第i类聚类的目标函数；K为聚类个数；X_j为第j个输入向量；C_i为第i个聚类中心(向量)，w_{ji}为权重(隶属度矩阵)。

首先，选择k个初始质心，每个点指派到最相近的质心，而指派到一个质心的点集为一个簇。然后，根据指派到簇的点，更新每个簇的质心。重复指派和更新步骤，直到簇不发生变化，或等价地，直到质心不发生变化。

K-means算法的步骤如下。

Step1：随机选取k个数据点C_i，$i=1$，2，…，k；并将之分别视为各聚类的初始中心。

Step2：决定各数据点所属的聚类，若数据点X_j判定属于第i聚类，则权重值$w_{ji}=1$，否则为0。

$$w_{ji}=\begin{cases}1;\ if \left\|X_j-C_i\right\|\leqslant\left\|X_j-C_m\right\|,\ \forall m\neq j\\0;\ otherwise\end{cases}\tag{5-6}$$

且满足：

$$\sum_{i=1}^{k}w_{ji}=1;\ \forall j=1,2,\cdots,n;\ \sum_{i=1}^{k}\sum_{j=1}^{n}w_{ji}=n\tag{5-7}$$

Step3：计算目标函数J，如果J保持不变，代表聚类结果已经稳定不变，则可结束此迭代方法；否则进入Step4。

Step4：更新聚类的中心点，回到Step2。

$$C_i=\frac{\sum_{j=1}^{n}w_{ji}X_j}{\sum_{j=1}^{n}w_{ji}}\tag{5-8}$$

K-means算法是一种硬分类方法，它把每个待辨识的对象严格地划分到某个类中，具有非此即彼的性质。而实际上遥感数据所反映的大多数地物覆盖在形态和类别方面存在着中介性，没有确定的边界来区分它们。因此，需要考虑各个像元属于各个类别的隶属度问题，对部分混合像元进行软划分，才能更好地区分不同的地物

类别[22]。

2. 模糊c-means算法

模糊c-means(FCMclust)算法[23]的目标函数定义如同K-means聚类算法，但其权重矩阵W不再是二元矩阵，而是应用了模糊理论的概念，使得每一输入向量不再仅归属于某一特定的聚类，而以其归属程度来表现属于各聚类的程度。

目标函数J为：

$$J = \sum_{i=1}^{K} J_i = \sum_{i=1}^{K} \left(\sum_{j=1}^{N} w_{ji}{}^{m} \left\| X_j - C_i \right\|^2 \right) \tag{5-9}$$

其中，X_j为数据点；C_i为聚类中心点；N为数据个数；K为聚类中心点个数；m为权重指数；w_{ji}为权重(隶属度矩阵)。

FCMclust算法的步骤如下。

Step1：设定分类个数k及初始权重矩阵，随机给定0～1之值，并满足权重总和为1，公式如下。

$$\sum_{i=1}^{K} w_{ji} = 1;\ \forall w_{ji} \in [0,\ 1];\ j = 1,\ 2,\ \cdots,\ N;\ 0 < \sum_{j=1}^{N} w_{ji} < N \tag{5-10}$$

Step2：计算聚类中心点。

$$C_i = \frac{\sum_{j=1}^{n} w_{ji}{}^{m} X_j}{\sum_{j=1}^{n} w_{ji}{}^{m}} \tag{5-11}$$

Step3：计算目标函数值，当目标函数值小于设定的容忍误差可结束迭代过程；否则执行Step4。

$$E(t) = \| J^{(t)} - J^{(t-1)} \| < \varepsilon \tag{5-12}$$

Step4：重新计算权重矩阵w，并回到Step2进行运算。

$$w_{ji}=\frac{1}{\sum_{s=1}^{K}\left(\frac{\|X_j-C_i\|}{\|X_j-C_s\|}\right)^{\frac{2}{m-1}}} \tag{5-13}$$

比之脆弱或硬分割方法，FCMclust能够保留初始图像的更多信息。然而，FCMclust的一个缺点是不考虑图像上下文中的任何空间信息，这使得它对噪声和其他人造图像非常敏感。另外，由于采用平方误差和准则，该方法仅适合于发现球形或类似球形分布的类别，而大多数多光谱遥感图像的散点图趋于椭球体分布[24]。因此，一些学者在FCM算法的基础上，通过修改准则函数，达到对不同形状分布样本的聚类[25]~[26]，例如适合椭球形分布样本聚类的Gustafson-Kessel算法。

3. 模糊Gustafson-Kessel聚类算法

Gustafson-Kessel(GKclust)算法是距离自适应动态聚类算法(Adaptive Distance Dynamic Clustering Algorithm)的模糊推广，它可以有效地搜索超椭球、平面或线型的数据类[27]。在GKclust算法中，n维数据空间中点x_k到聚类中心v_i的距离是一个平方内积距离范数：

$$D^2_{ikMi}=(x_k-v_i)^T M_i(x_k-v_i);\ 1\leqslant i\leqslant c,\ 1\leqslant k\leqslant N \tag{5-14}$$

其中，$M_i=\det(F_i)^{1/n}F_i^{-1}$，$F_i$是第$i$个聚类中心的协方差矩阵，为正定对称矩阵。将数据集$\{x_1, x_2, \cdots, x_N\}$划分为$c$个模糊类是通过最小化目标函数来完成的。公式如下：

$$J(X;U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m\|x_k-v_i\|^2_{M_i} \tag{5-15}$$

其中，$U=[u_{ik}]$是数据集的模糊划分矩阵，且满足：

$$\sum_{i=1}^{c}u_{ik}=1;\ 1\leqslant k\leqslant N,\ u_{ik}\in[0, 1] \tag{5-16}$$

$m\in[1，\infty]$为一个加权指数，决定着所有分类的模糊程度。Lagrange乘子λ_k可以将目标函数式(5-15)及其约束式(5-16)转化为新的目标函数式(5-17)：

$$\bar{J}(X;U,V,\lambda)=\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ik})^m D_{ikm_i}^2+\sum_{k=1}^{N}\lambda_k[\sum_{i=1}^{c}u_{ik}-1] \tag{5-17}$$

设$D^2_{ikMi}>0$，$\forall i$，k及$m>1$，另J关于U，V和λ的梯度为0，则可求得使式(5-15)取极小值的两个必要条件：

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}(D_{ikA_i}(x_k,v_i)/D_{jk}(x_k,v_j))^{2/(m-1)}}；\ 1\leqslant i\leqslant c,\ 1\leqslant k\leqslant N \tag{5-18}$$

$$v_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^m x_k}{\sum_{k=1}^{N}(\mu_{ik})^m}；\ 1\leqslant i\leqslant c \tag{5-19}$$

给定数据集$\{x_k|k=1，2，\cdots，N\}$，选择分类数目$1<c<N$，加权指数$m>1$和终止容许误差$\varepsilon>0$。随机初始化模糊划分矩阵U，并使之满足式(5-16)，设l为迭代次数，算法步骤如下。

Step1：计算聚类中心点。

$$v_i^l=\frac{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^{N}(\mu_{ik}^{(l-1)})^m}；\ 1\leqslant i\leqslant c \tag{5-20}$$

Step2：计算类协方差矩阵。

$$F_i=\frac{\sum_{k=1}^{N}(u_{ik}^{(l-1)})^m(x_k-v_i^{(l)})(x_k-v_i^{(l)})^T}{\sum_{k=1}^{N}(u_{ik}^{(l-1)})^m}；\ 1\leqslant i\leqslant c \tag{5-21}$$

Step3：计算距离。

$$M_i=\det(F_i)^{1/n}F_i^{-1} \tag{5-22}$$

$$D_{ikMi}^2=(x_k-v_i^{(l)})^T M_i(x_k-v_i^{(l)}\ );\ 1\leqslant i\leqslant c,\ 1\leqslant k\leqslant N \tag{5-23}$$

Step4：更新模糊划分矩阵。

$$\mu_{ik}^{(l)}=\frac{1}{\sum_{j=1}^{c}(D_{ikA_i}(x_k,v_i)/D_{jk}(x_k,v_j))^{2/(m-1)}};\ 1\leqslant i\leqslant c,\ 1\leqslant k\leqslant N \tag{5-24}$$

Step5：判别迭代条件。

如果满足下式：

$$||U^{(l)}-U^{(l-1)}||<\varepsilon \tag{5-25}$$

结束迭代过程。

5.3.3　聚类有效性验证

评价聚类结果优劣的过程，称为聚类的有效性验证。一般来讲，使类内距离极小化而类间距离最大化的聚类是最优聚类。如下给出不同的有效评估方法来判别不同算法的聚类效果。其中以N代表数据个数，c代表聚类个数，c_i代表第i个聚类，v_i代表第i个聚类的中心点，u_{ij}表示点x_i属于c_j的隶属度。

1. 划分系数PC[28]

Bezdek对模糊聚类设计划分系数(Partition Coeficient)定义如下：

$$PC(c)=\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2 \tag{5-26}$$

该系数用于评判分类簇之间的分离程度。在分类簇数目相同的情况下，PC值越接近1，分类效果越好。其缺点在于该指标值和隶属度有关，和数据的其他属性缺乏联系，并随c的增加单调下降。

2. 分类熵CE[28]

分类熵CE(Classification Entropy)定义如下：

$$CE(c)=-\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log(\mu_{ij}) \tag{5-27}$$

用于计算分类簇的模糊度。在分类簇数目相同的情况下，CE值越小，分类效果越好。

3. Xie and Beni's Index 指标XB[29]

反应分类簇内的紧致性和分类簇间的分离性。其定义如下：

$$XB(c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{m}\left\|x_j-v_i\right\|^2}{N\min_{i,j}\left\|x_j-v_i\right\|^2} \tag{5-28}$$

在分类簇数目相同的情况下，XB值越小，分类效果越好。

5.4 一种新的自训练半监督支持向量机分类模型构建

为了克服自训练半监算法“错误累积”不足，本书提出了一种新的自训练半监督支持向量机分类模型(简称PSVM)，主要有如下特点：

首先，从分类器的构造角度，利用第4章介绍的自适应变异粒子群算法(AMPSO)优化SVM分类器参数构建AMPSO-SVM(以下简称PSVM)分类模型作为基分类器，使用PSVM对可靠未标注样本集合中的元素进行反复迭代，扩大标记样本数目。

其次，未标记样本池选择过程中，引入模糊聚类算法对未标签样本产生模糊隶属度函数，利用模糊隶属度函数将最接近样本的有效无标签样本作为标注对象，远离标签的无效样本值，以控制错误信息的输入。

5.4.1　未标记样本的选择依据

PS3VM模型首先利用初始训练集训练PSVM得到初始分类器，采用Gkclust模糊聚类算法对未标记样本进行模糊聚类产生模糊隶属度μ_i，然后根据如下公式选择高于$\overline{U}$的作为未标签样本的候选集合，进而使分类间隔最优化，如图5-2所示。

$$\overline{U}=\sum_{i=1}^{t}\mu_i\Big/t \tag{5-29}$$

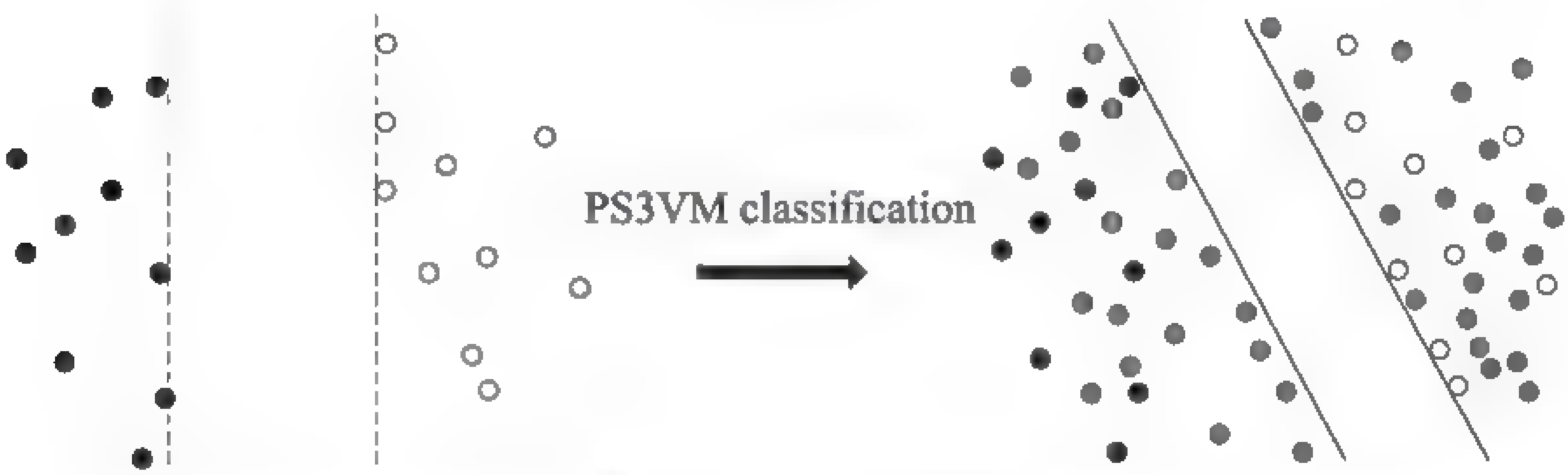


图5-2　PS3VM分类

注：图5-2中，黑色和白色圆圈分别代表标签样本。左侧代表只有标签样本参与分类时的分类间隔(虚线)；右侧表示标签样本与新标签样本(灰色圆圈)共同参与分类时的分类间隔(实线)。

5.4.2　基于GKclust的自训练半监督支持向量机设计流程

为了更清楚地说明所提出的学习技术，有必要对如下概念进行说明。

标签样本集：$L=\{L_1, L_2, \cdots, L_l\}$；

无标签样本集：$M=\{M_1, M_2, \cdots, M_n\}$；

当前新样本集：$T=\{T_1, T_2, \cdots, T_k\}$，即初始样本集合与新标

签样本集合的并集；

聚类中心点集：$V=\{V_1, V_2, \cdots, V_c\}$。

PS3VM分类模型首先利用Gkclust模糊聚类算法对初始标签样本点L进行非监督聚类产生c个类别的聚类中心T；然后以T为初始聚类中心，利用Gkclust对无标签样本点M进行非监督聚类产生c个类簇和所有无标签样本点的模糊隶属度函数μ_i，在各个类簇中，将距离聚类中心较近的点(高于$\overline{U}$)作为未标签样本的候选集合N；利用PSVM模型，同时设定一个阈值τ对候选集进行样本标注形成集合ψ；接下来将ψ增加到L中，并在M中将ψ删除，如此迭代，直至M为空，流程图如5-3所示。

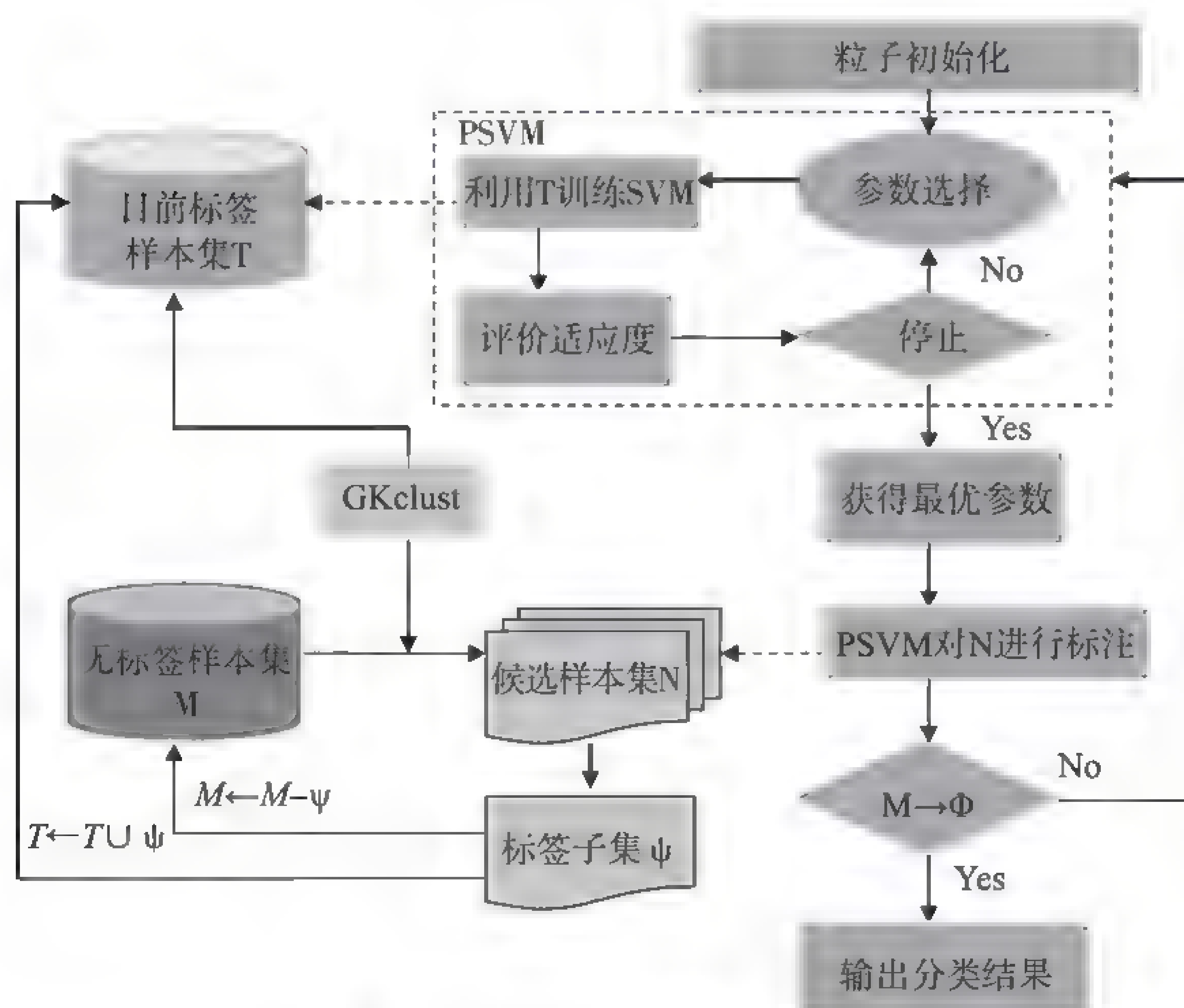


图5-3　PS3VM分类模型算法流程

5.4.3 基于GKclust的自训练半监督支持向量机算法

Step1：初始化标签样本集T=L，无标签样本集M，$\tau=\tau_0$；

Step2：当$M\neq\Phi$执行如下操作；

Step3：利用标签集训练SVM，并利用AMPSO进行参数优化，构建初始分类器；

Step4：在集合T中利用Gkclust模糊聚类算法根据公式(5-19)产生聚类中心V；

Step5：以V为初始聚类中心，在无标签集合中根据公式(5-18)产生无标签样本的模糊隶属度函数值；

Step6：将隶属度高(高于$\overline{U}$)的样本点组成候选集合N；

Step7：利用PSVM对N进行标注；

Step8：基于τ产生标签子集ψ；

Step9：更新标签集 $T\leftarrow T\cup\psi$；

Step10：更新无标签集 $M\leftarrow M-\psi$；

Step11：如果$\psi=\Phi$，降低τ的值；

Step12：判断循环是否结束；

Step13：利用T再次训练PSVM。

5.5 实验结果与分析

在本节中，为了评价所提算法的性能，首先将研究区Landsat-5 TM遥感数字化，然后分别针对TM数字数据和TM影像数据进行分类

实验。在第一个试验里，Gkclust聚类算法与其他相关算法，如fuzzy c-means(FCMclust)和K-means，在聚类有效指数、聚类精度方面进行比较；评价在半监督学习过程中，无标签样本的加入数量与分类精度的关系，以及分类参数的变化情况。第二个实验，利用无标签样本与标签样本的合理比例、最优分类参数构建优化模型，并将其应用于影像集的分类实验。

5.5.1 遥感影像数字化

本文选择2009年9月30日获取行列号115-30多光谱Landsat-5 TM遥感影像(30米空间分辨率，UTM投影)。根据第3章的介绍，本试验数字影像包括8个特征，分别是TM图像的6个波段(1～5，7)、PCA的第一主分量、植被指数(NDVI)。实验区按第3章介绍的二级分类系统分为6个土地利用类型，即落叶针叶林(DCF)、常绿针叶林(ECF)、落叶阔叶林(DBLF)、旱地(FL)、居住地(RL)、内陆水体(WT)。为了保证每个类别数据的变化性和代表性，数字集采用随机像素的选择策略[30]。然后将数字集分成两个子集，一部分用于训练，另一部分用于测试。具体的土地覆盖类型及训练样本、测试样本的数量如表5-1所示，图5-4显示了1 874个样本点的分布情况。

表5-1 类别及样本数量

类别缩写	类别名称	训练集数量	测试集数量
DCF	落叶针叶林	150	149
FL	旱地	151	151
RL	居住地	155	234
ECF	常绿针叶林	116	102

(续表)

类别缩写	类别名称	训练集数量	测试集数量
DBLF	落叶阔叶林	201	217
WT	内陆水体	121	127
6	6	894	980

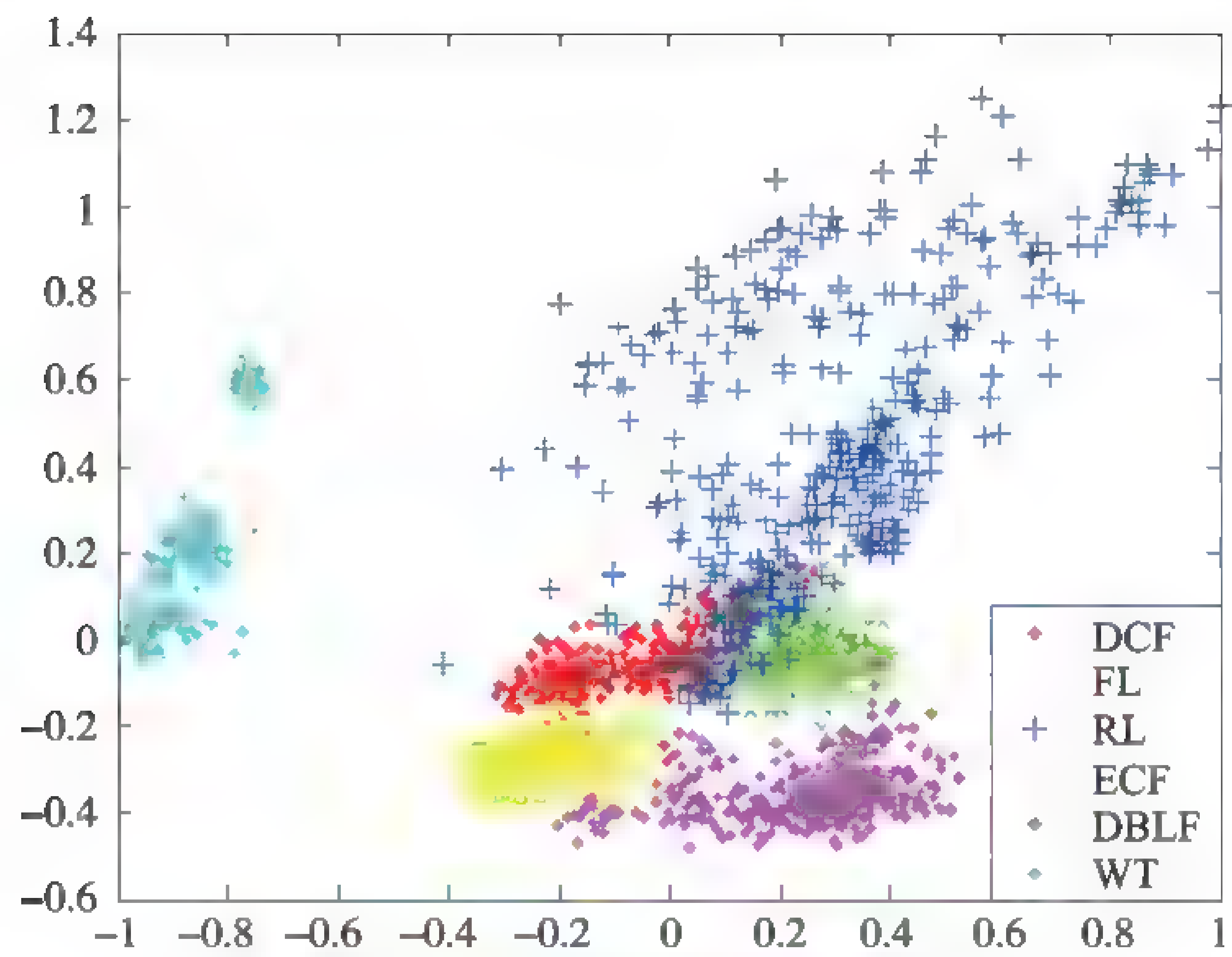


图5-4　1 874个样本点的分布情况

5.5.2　参数设置

为了达到较好的收敛，参数c和γ的取值范围分别为[0.1，100]和[0.01，1 000]，其他相关参数如表5-2所示，此外，阈值τ的初始值为0.7。

表5-2　参数取值

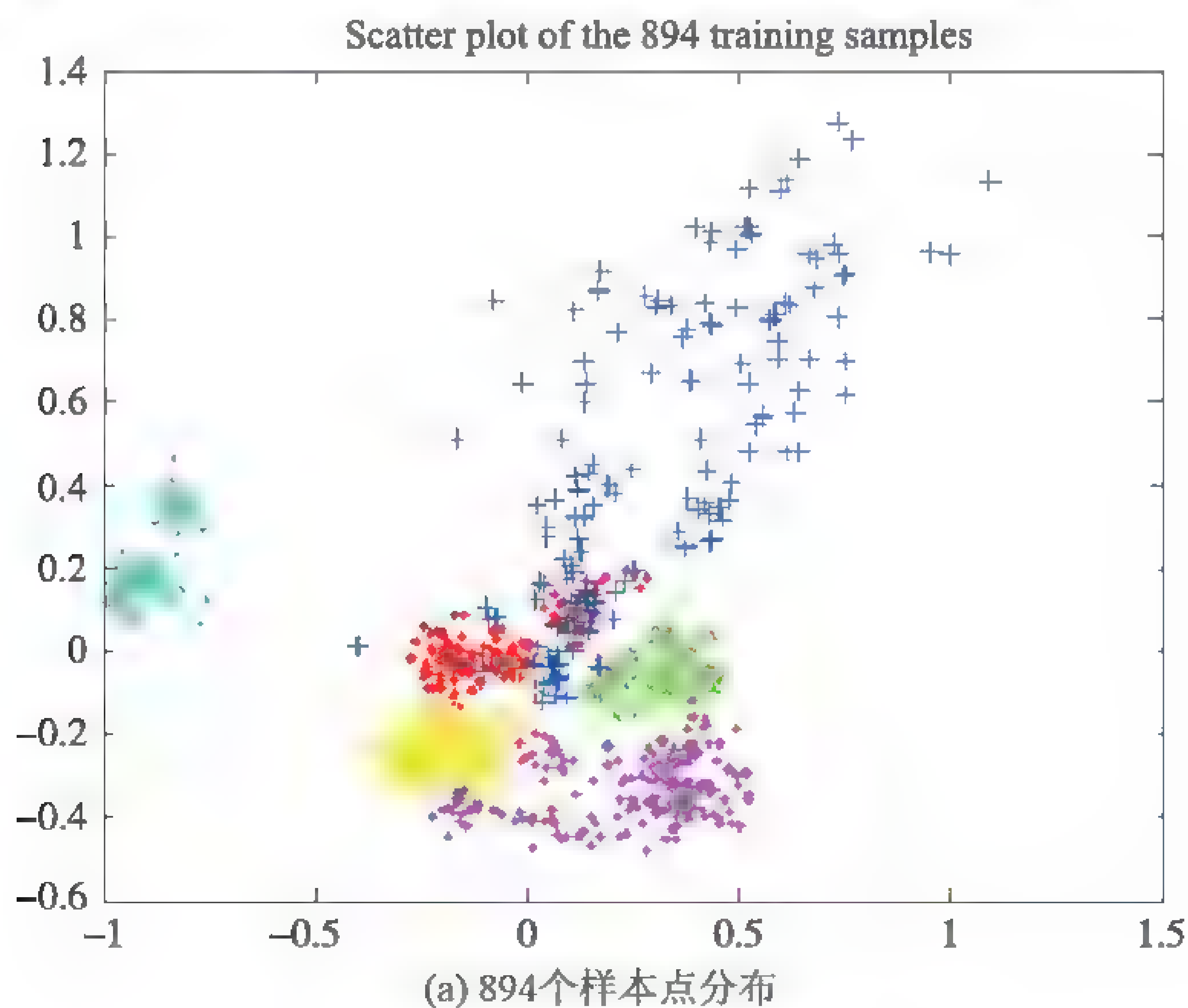
Coefficient	Values
c_1	1.5
c_2	1.7
maxgen	200

(续表)

Coefficient	Values
sizepop	20
cmax	100
cmin	0.1
γmax	1 000
γmin	0.01

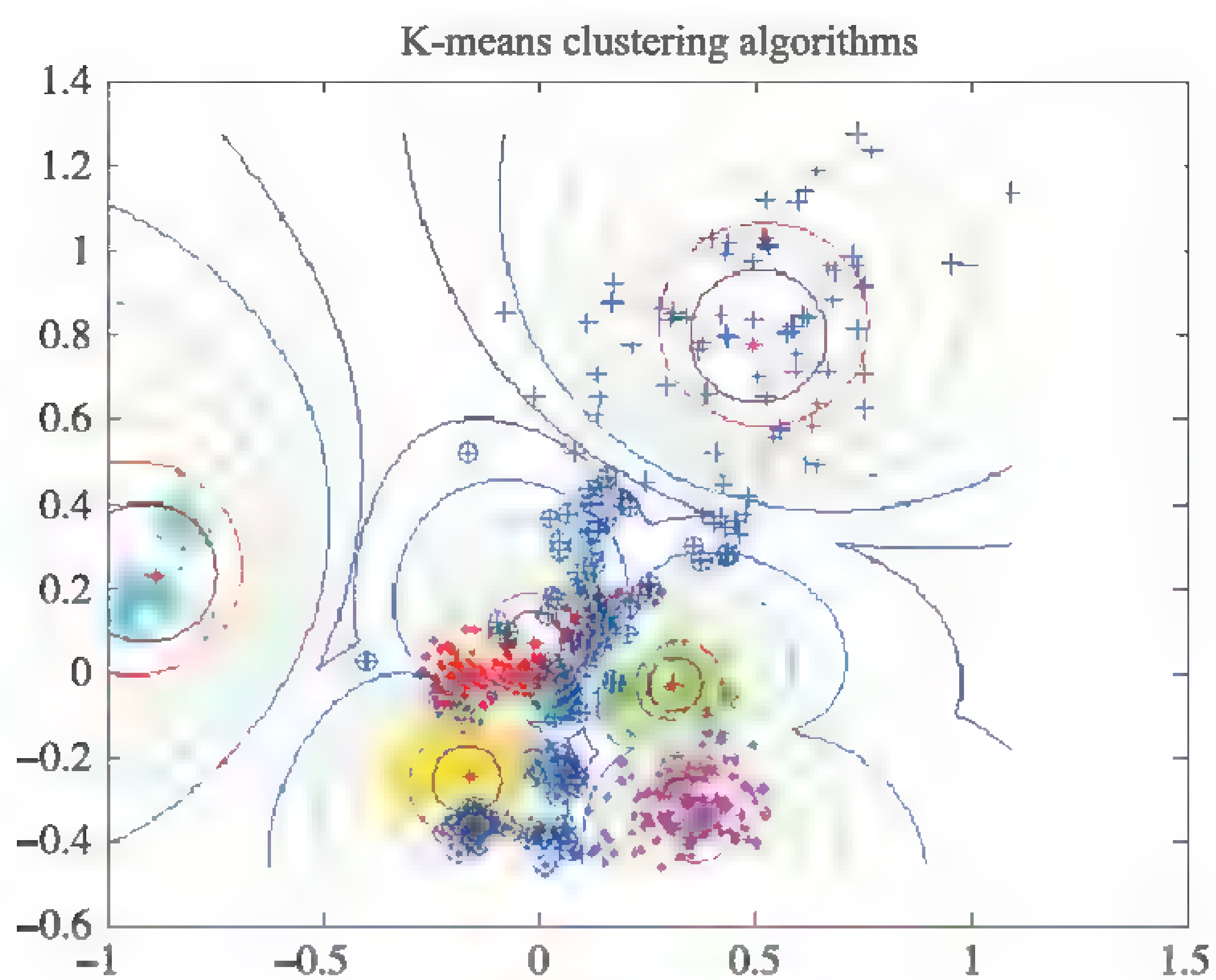
5.5.3 模糊聚类算法的比较

为了达到比较的目的，本书分别用GKclust、FCMclust、K-means模糊算法针对同一训练集进行聚类，并将聚类结果在聚类有效性指数(划分系数PC、分类熵CE、XB指标)和聚类精度等方面进行比较，如图5-5所示。其中，图5-5(a)为894训练样本点的分布情况，图5-5(c)～(d)分别为3种算法对894训练样本的聚类结果。

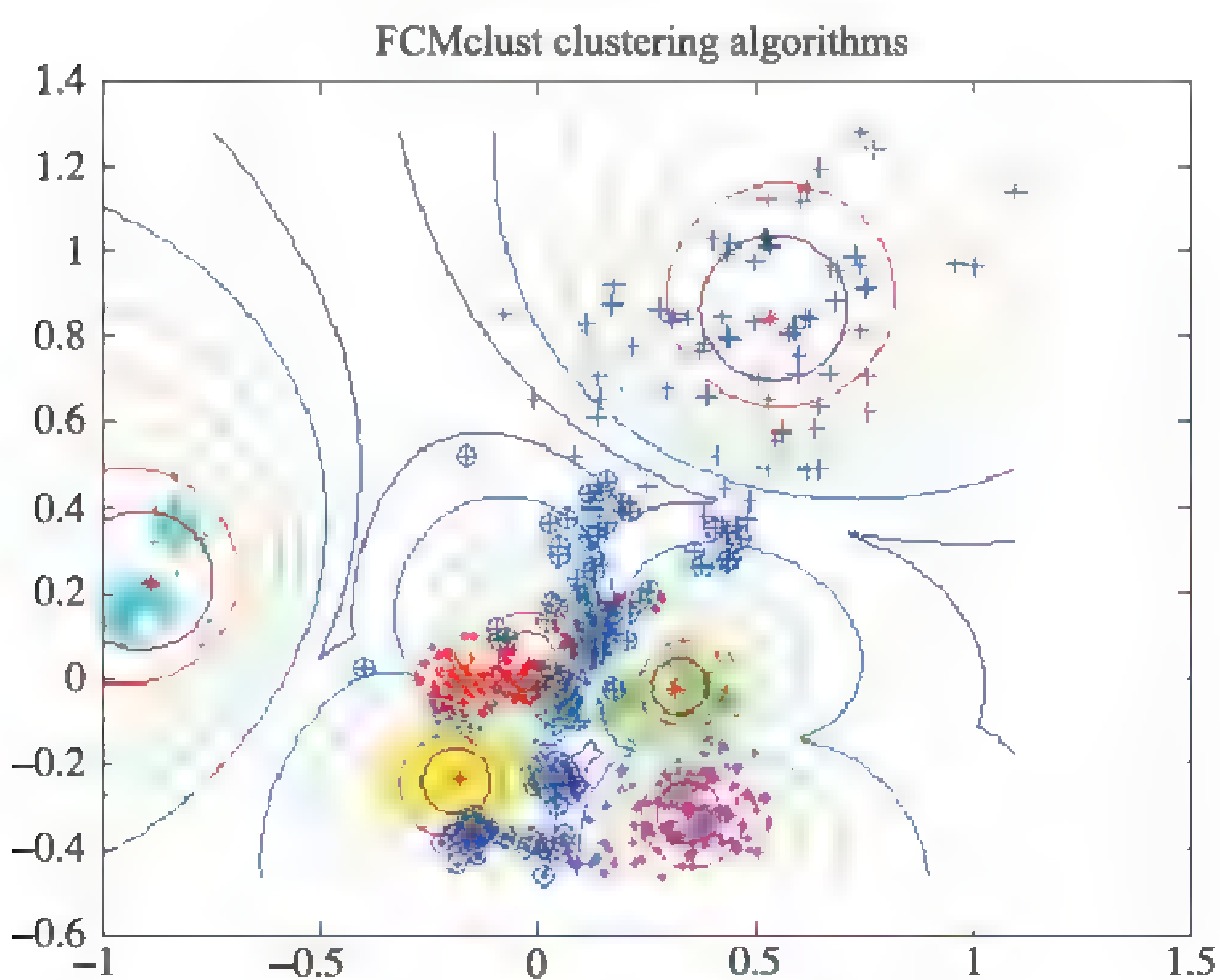


(a) 894个样本点分布

图5-5 模糊聚类算法的比较

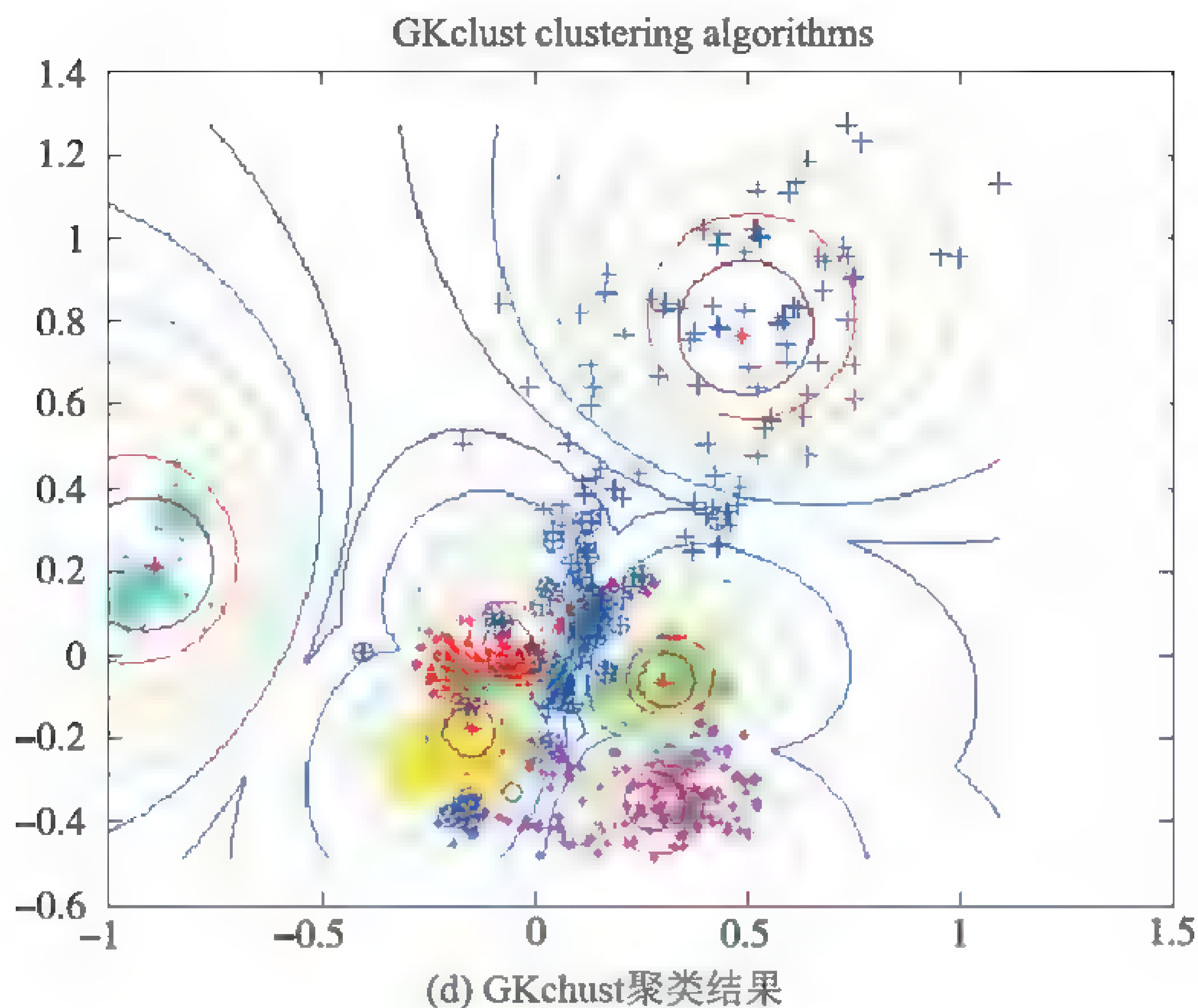


(b) K-means聚类结果



(c) FCMclust聚类结果

图5-5　模糊聚类算法的比较(续)



(d) GKchust聚类结果

图5-5 模糊聚类算法的比较(续)

其中，'*'代表类簇的聚类中心，错分样本点用'o'标识。为了达到可视化效果，二维空间中利用等高线图描述数据分布。从图中可以看到6个地物类别分别对应6个椭圆形集群的叠加，由于GKclust算法通过自适应距离拓展了FCMclust算法使得椭球形被拉长，可以明显地看到GKclust算法中错分类别的点集明显少于FCMclust和K-means方法。为了更好地比较3种聚类算法对该研究区数字集聚类的不同效果，表5-3也列出不同聚类算法所产生的有效指数及聚类精度(Accuracy，ACC)。从表5-3不难发现，对于硬聚类算法K-means，PC和CE值是无效的。对比3个聚类指数，GKclust算法针对这一数据集产生的结果最好。由GKclust产生的聚类精度也比FCMclust和K-means高出5%以上。因此，本书利用GKclust对未标记

样本进行优选，以提高标注效率。

表5-3　聚类有效性指数值

Method	PC	CE	XB	ACC
GKclust	0.690 5	0.610 1	1.862 7	91.05
FCMclust	0.641 8	0.777 6	2.966 6	85.23
Kmeans	1	NaN	4.274 5	85.01

5.5.4　无标签样本的参与比例

在标签样本集有限的情况下，无标签样本的参与可以改善分类性能。然而，在许多情况下，过多或无效的无标签样本可能会降低分类性能[31]。为了观察未标记样本的参与是否对分类精度的提高有帮助，以及如何搭配未标记样本和已标记样本的数目，才能使分类的投入最少、效率最高。本实验分别选取了不同标签样本和无标签样本组合。具体为分别从每个类别独立选取25个、50个、75个、100个已标记样本(150，300，450，600)，以及整个训练样本(894)，共存储为5个已知标签样本集，并随机选取500个、1 000个、2 000个、3 000个和5 000个未标记样本，存储为5个未知标签样本集，进行半监督分类实验。依次从5个标记样本集和5个未标记样本集中选取两个样本集组合进行分类实验，共有25个组合。

25个组合利用PS3VM，S3VM(FCMclust模糊聚类的半监督PSVM简称)半监督分类结果以及利用PSVM对5个已知标签样本集监督分类结果如表5-4所示。同时图5-6(a)～(e)列出3种方法分类精度随标签样本数量变化的曲线；图5-6(f)显示50组实验得出的较好精度下标签样本和无标签样本的比例图。从表5-4和图5-6观察可得出如下

重要结论：

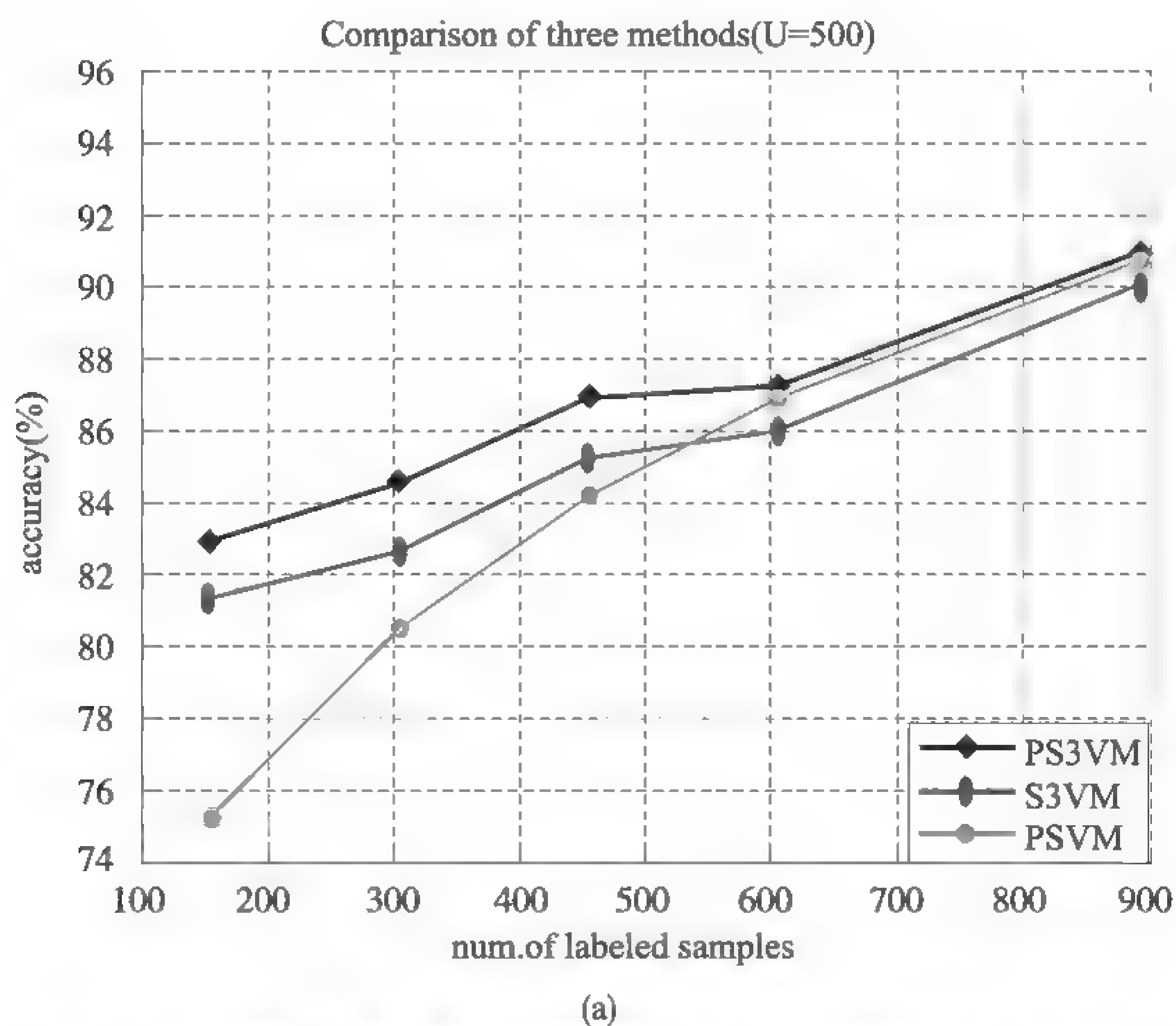
首先，未标记样本参与的半监督分类方法可有效提高已标记样本代表性不好时的分类精度(如图5-6(a)～(e)所示)，特别是对较小训练样本集效果更显著。

其次，随着已标记分类样本的增加，未标记样本的作用越来越小。从表5-4分类精度较好的结果可知，对于L=150，U=500这一组合(其中L代表标签标本数，U代表无标签样本数)，PS3VM所获精度82.92%，而PSVM所获精度为75.23%，即PS3VM分类精度高于PSVM分类精度7.69%；随着标签样本数量的增多，对于L=894，U=3 000这一组合，PS3VM所获精度为95.10%，而PSVM所获精度90.81%，相比之下，PS3VM分类精度高于PSVM分类精度4.29%。

最后，从实验结果可以看出，冗余的无标签样本可能会降低分类精度，例如，对于组合L=150，U=5 000，S3VM所获得的分类精度甚至要比仅使用标签样本(L=150)的PSVM获得的分类精度还低，也就是说，要想利用半监督分类获得较小的分类误差，标签样本和无标签样本应该满足一定的比例。由表5-4可以看出，对于PS3VM半监督分类模型，如下组合分类结果较好，L=150，U=500；L=300，U=1 000；L=450，U=1 000；L=600，U=2 000；L=894，U=3 000；对于S3VM半监督分类模型，如下组合分类结果较好，L=150，U=500；L=300，U=1000；L=450，U=2 000；L=600，U=3 000；L=894，U=3 000产生的分类精度较高，按照规律发现已标记样本和未标记样本之间保持的合适比例关系大约在0.3(1:3)左右(如图5-6(f)所示)。

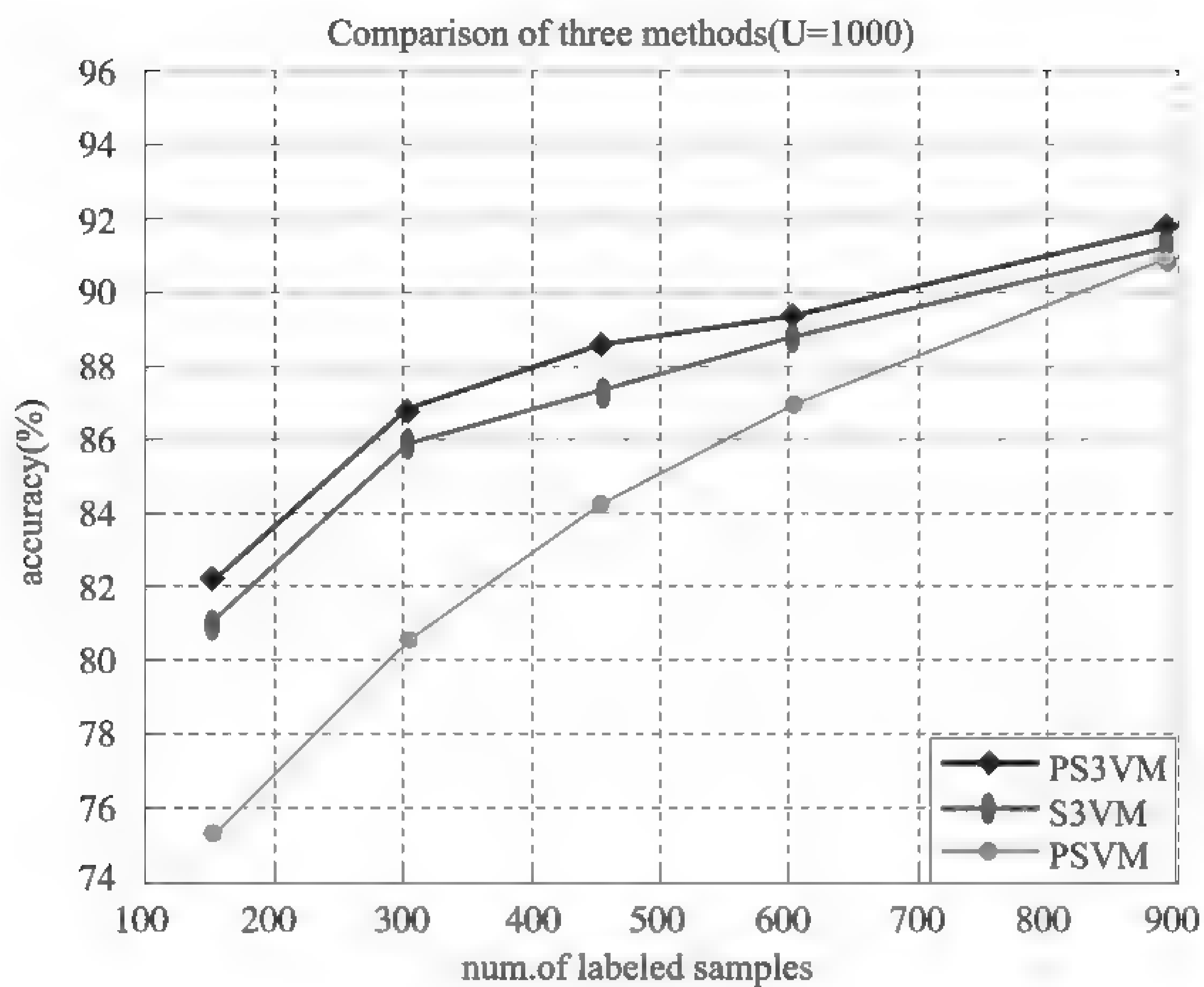
表5-4　在标签样本数量变化时产生的分类精度

Method	U	L=150	L=300	L=450	L=600	L=894
PS3VM	500	82.92	84.56	86.98	87.25	90.98
	1 000	82.23	86.71	88.45	89.26	91.67
	2 000	81.81	86.19	89.94	91.64	92.92
	3 000	80.96	85.89	88.12	91.27	95.10
	5 000	77.56	84.45	87.89	88.92	93.61
S3VM	500	81.34	82.67	85.24	85.98	90.03
	1 000	80.87	85.89	87.24	88.67	91.04
	2 000	79.29	84.74	88.55	90.21	92.30
	3 000	78.05	84.08	87.30	90.78	93.06
	5 000	75.02	82.21	86.88	87.89	91.61
PSVM	0	75.23	80.50	84.20	86.92	90.81

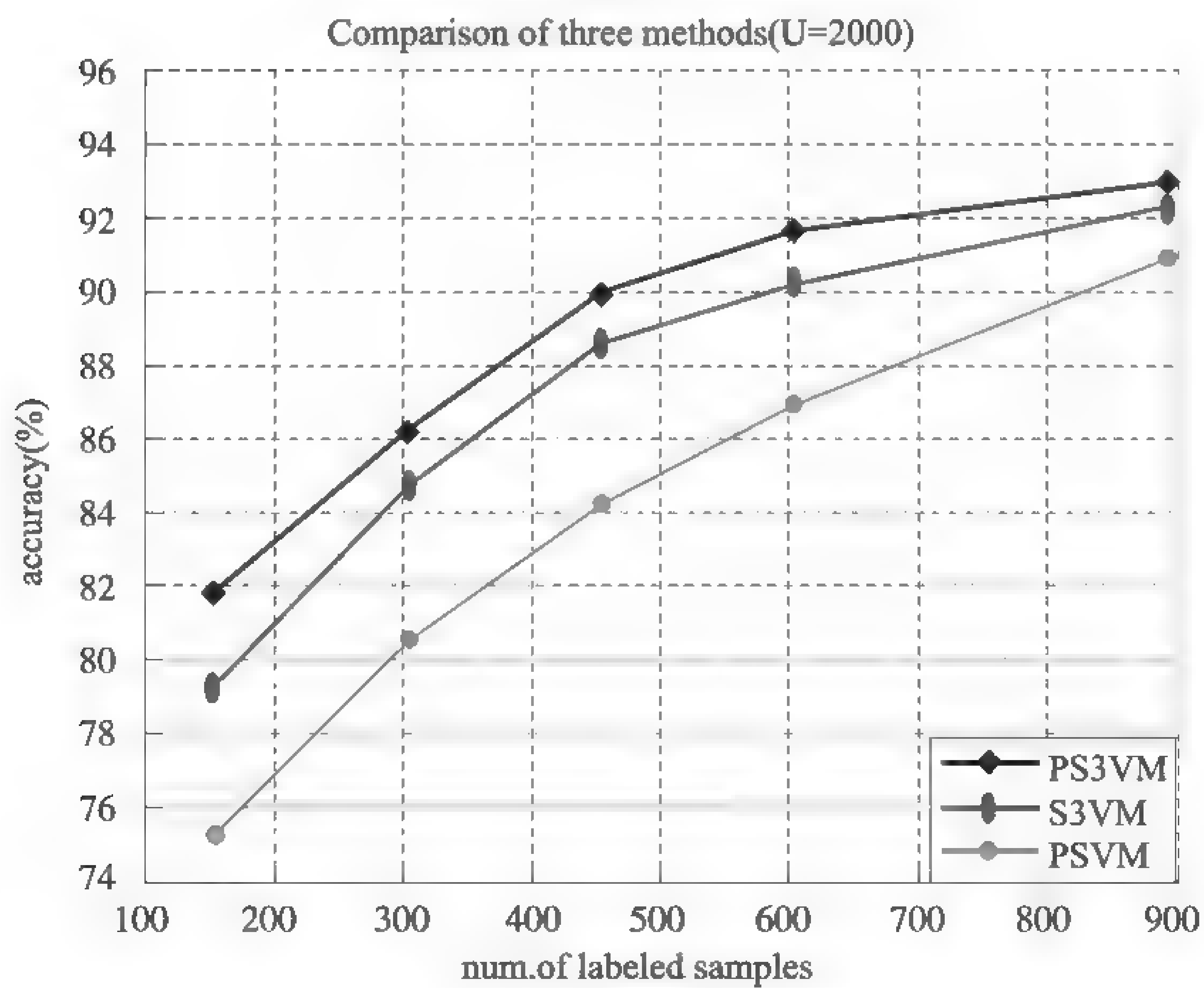


(a)

图5-6　三种方法分类、精度随标签样本及无标签样本数量变化的曲线

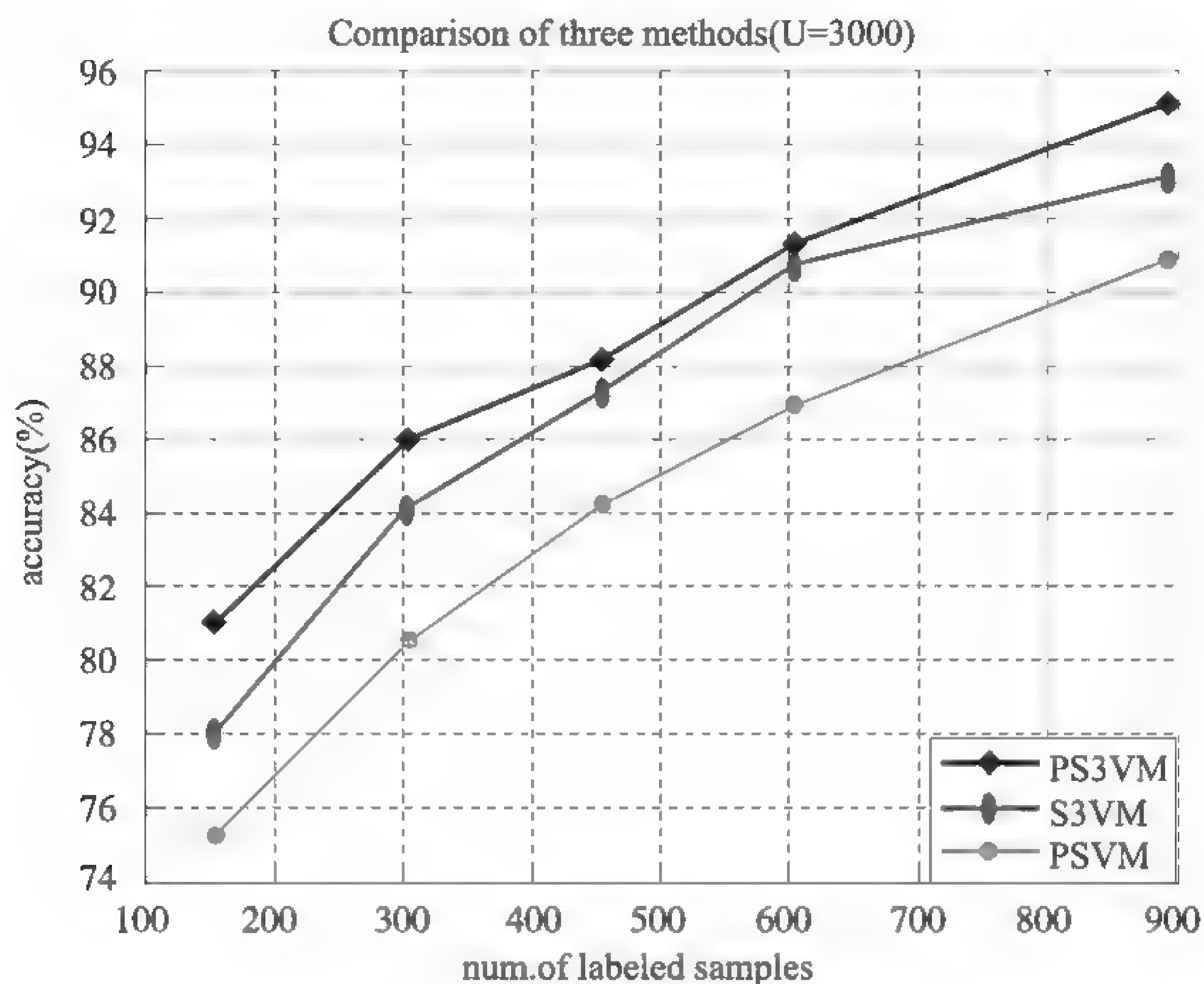


(b)

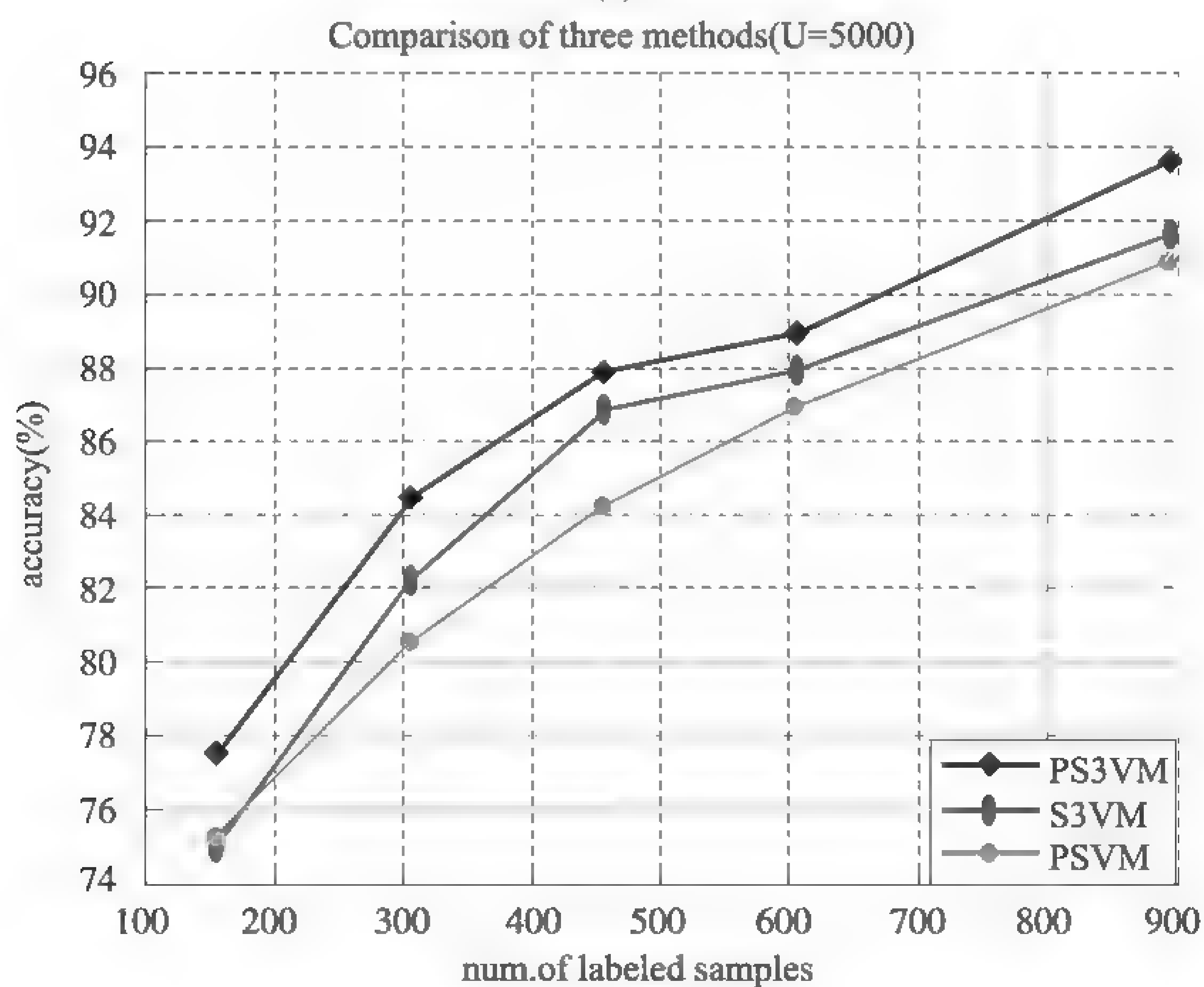


(c)

图5-6 三种方法分类、精度随标签样本及无标签样本数量变化的曲线(续)

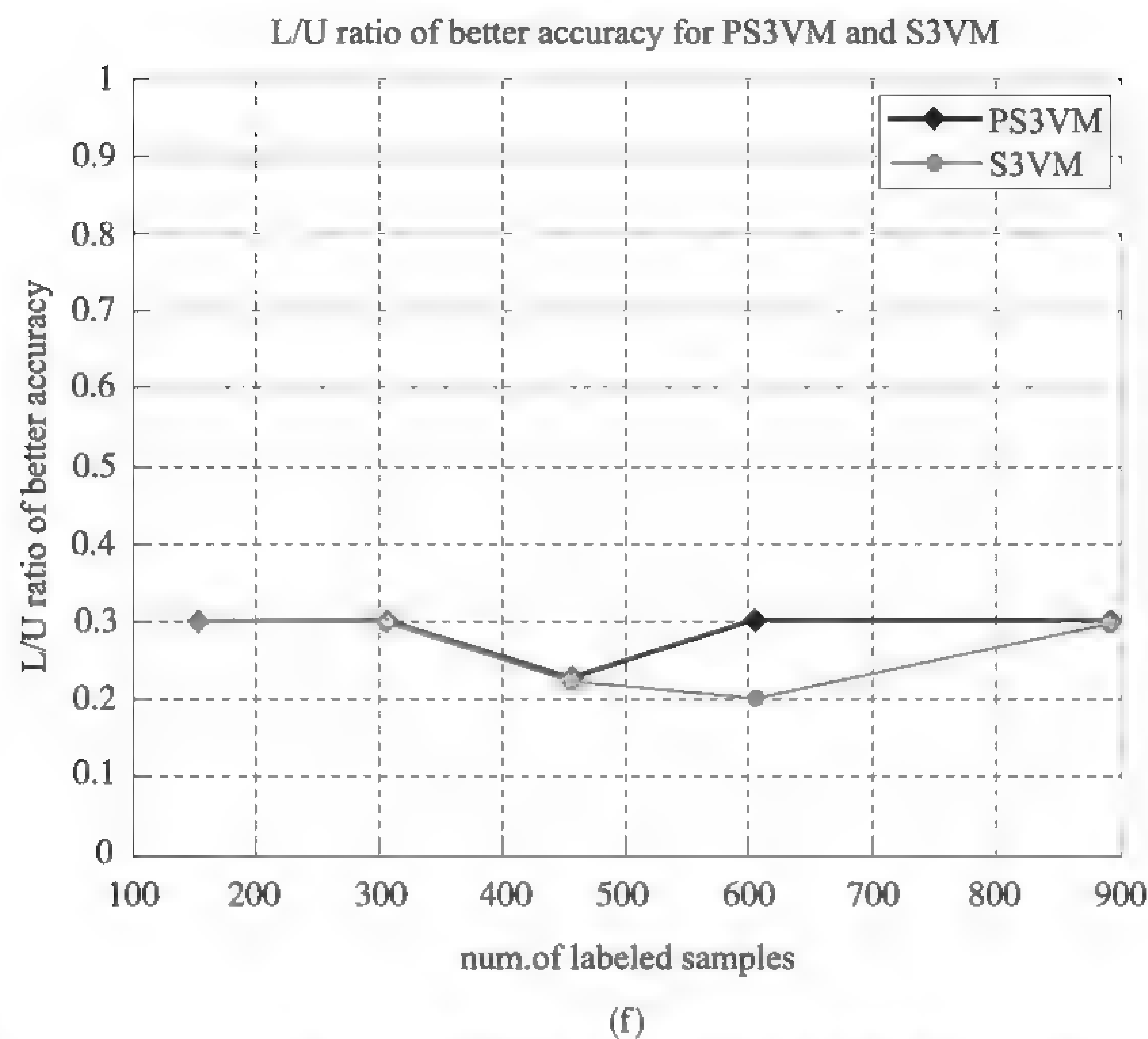


(d)



(e)

图5-6　三种方法分类、精度随标签样本及无标签样本数量变化的曲线(续)



(f)

图5-6　三种方法分类、精度随标签样本及无标签样本数量变化的曲线(续)

注：图5-6 其中(a)-(e)为3种方法随标签样本及无标签样本变化曲线；(f)表示PS3VM与S3VM所产生较好精度的标签和无标签样本比例。

表5-5　三种方法分类参数、分类精度、kappa系数值的比较

Method	$-c$	$-\gamma$	kappa	Overall accuracy		DCF	FL	RL	ECF	DBLF	WT
PS3VM	2.66	17.634 0	0.940 2	95.10	DCF	139	0	4	6	0	0
					FL	0	140	3	1	7	0
					RL	5	2	227	0	0	0
					ECF	6	0	0	91	5	0
					DBLF	0	5	0	4	208	0
					WT	0	0	0	0	0	127

(续表)

Method	−c	−γ	kappa	Overall accuracy		DCF	FL	RL	ECF	DBLF	WT
S3VM	35.63	28.106 4	0.915 4	93.06	DCF	133	1	7	8	0	0
					FL	0	137	5	0	9	0
					RL	8	3	221	2	0	0
					ECF	7	0	1	92	2	0
					DBLF	1	7	0	5	203	1
					WT	0	0	0	0	1	126
PSVM	22.21	8.634 0	0.888 0	90.81	DCF	129	0	9	10	1	0
					FL	0	133	4	2	12	0
					RL	9	4	217	3	1	0
					ECF	7	1	0	90	4	0
					DBLF	2	9	1	7	196	2
					WT	0	0	0	0	2	125

根据上述实验可知，对于894个标签样本，3 000个无标签样本参与的组合，利用PS3VM产生的分类结果最好。为了比较，针对这个组合分别采用S3VM和PSVM分类器进行分类，其结果如表5-5所示，除了分类精度值，还包括参数c、γ、Kappa系数的取值以及分类混淆矩阵。从表5-5可以发现，专著所提出的PS3VM方法相比其他两种方法性能更优。

5.5.5　土地覆盖遥感图像分类

利用遥感影像的数字集，有效地测试了PS3VM的性能。在本节，将PS3VM应用于覆盖研究区域115-30TM影像子集的分类并产生分类专题图。按照已知标签样本和无标签样本的取值比例(0.3)，针对整个测试集1 874个已知样本点，随机从影像上搜集6 000个未知标签点进行半监督分类。分类结果如图5-7所示。

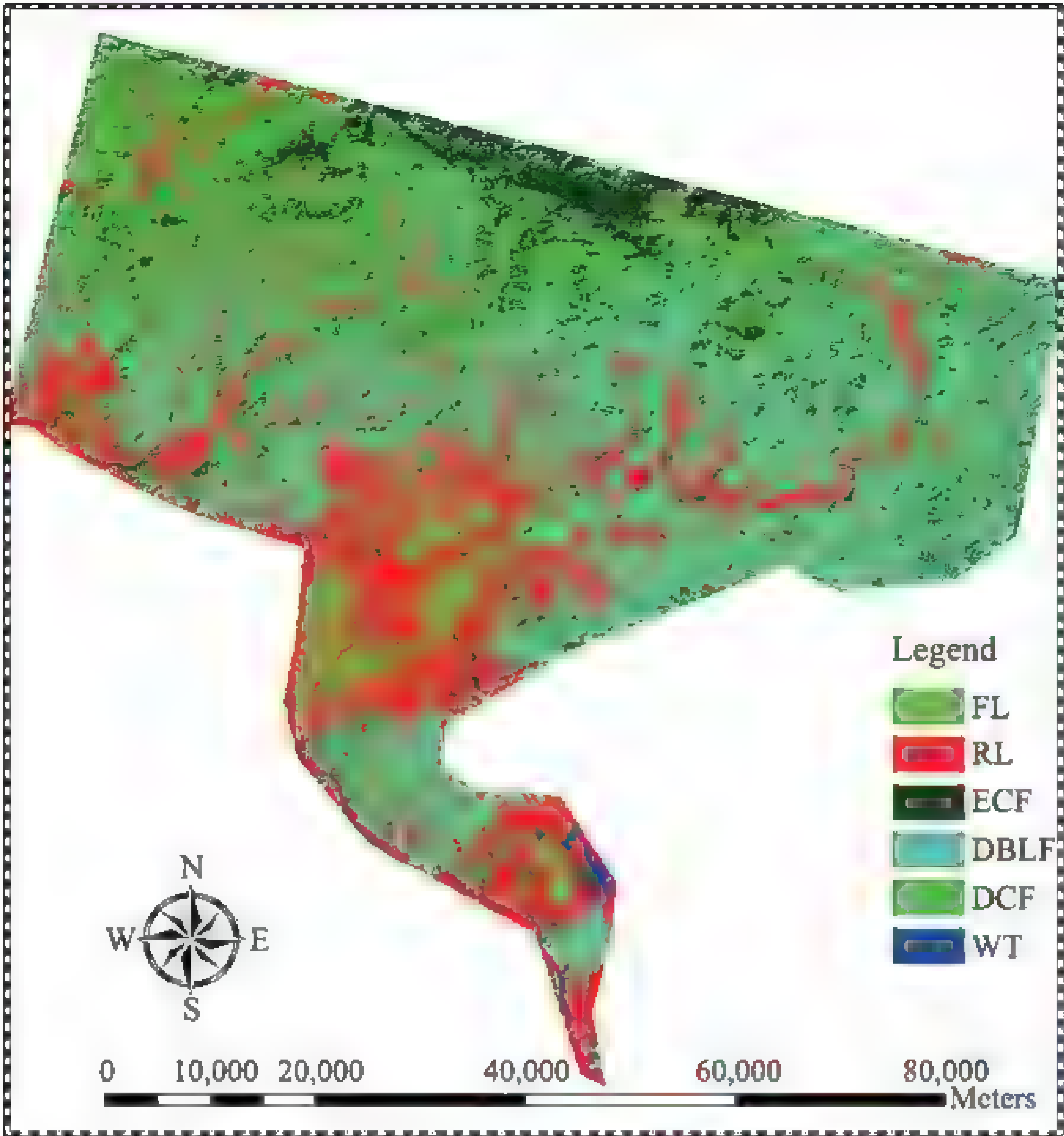


图5-7 利用PS3VM对TM影像分类结果图

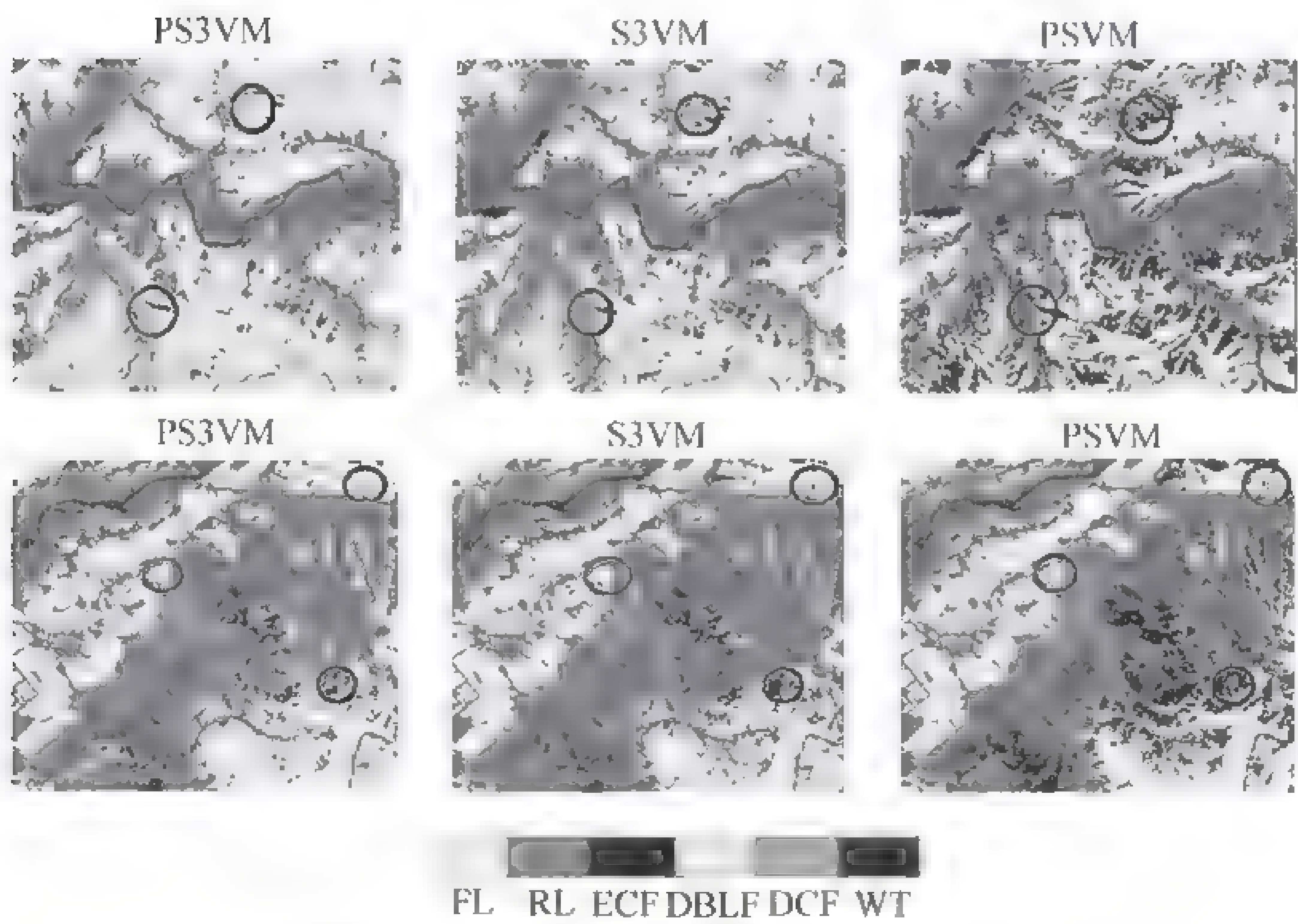


图5-8 三种方法产生的子专题分类图

注：图中圆圈标出的是3种方法分类差别显著位置。

从图中可以得出如下结论：首先，研究区域的主要土地覆盖类型为林地；其次，可以看出森林正面临着快速城市化的威胁，为了有效地保护森林资源，有必要对此研究区进行动态监测。

此外，为了观察3种算法对遥感影像数据分类精度的差异，专著从分类专题图中抽取几个典型区域用于比较，其分类专题的子集图如图5-8所示。3种算法明显的分类差异用黑色的圆圈标识，从左到右依次采用PS3VM、S3VM、SVM分类方法。从第一列能够看到，PS3VM能较好地解决针叶林和阔叶林混合像素的分类。第二列，PS3VM方法在耕地和城市用地，及林地之间的分类精度也明显高于其他两种方法。从子专题图中也不难发现，利用SVM方法很难解决常绿针叶林与阴影之间的混合像素分类识别问题。

综上可知，PS3VM在遥感影像数据上的分类也明显优于其他两种方法，无标签的样本有效加入确实能提高影像的分类精度。

5.6　本章小结

在遥感影像分类领域，构建出高精度、高性能的分类器通常不是一件容易的事情，因为任何分类模型的创建都离不开充足的、准确的训练样本，训练样本的获取不但费时、费力，而且还可能由于训练样本人为主观选择而产生已知样本的不准确性。半监督学习很好地解决了因样本不足而造成分类器性能低下这一问题。本文提出一种新的利用自训练算法融合启发式算法的半监督SVM分类模型，并且在标签样本的标注过程中，引入了GKclust模糊聚类算法，主要

目的在于通过无标签样本的合理利用以及准确的分类参数获得分类性能高的分类模型。并且将所提到的半监督学习模型分别应用于遥感的数字数据和影像数据的分类实验，可以得出如下结论：

(1) 无标签样本参与能够提高分类器的分类性能，但其数量要满足一定的比例；

(2) 针对本书所提供的数据集分布特点，GKclust算法相比于FCMclust、K-means算法确实有较高的聚类性能；

(3) 从分类比较结果可以看出，PS3VM能够有效克服自训练半监督模型中“错误累积”这一弊端，能够明显提高数据集的分类精度。

参考文献

[1] Serge A., Ludovic R., Yannick C., Alain B.. *A Fuzzy-possibilistic Scheme of Study for Objects with Indeterminate Boundaries:Application to French Polynesian Reefscapes*[J]. IEEE Transaction on Geoscience and Remote Sensing, 2000, 38(1): 257-270.

[2] Zadeh L.A.. *Fuzzy Sets*[J]. Infromation and Control, 1965, 8: 338-353.

[3] Wang F.. *Improving Remote Sensing Image Analysis through Fuzzy Information Representation*[J]. Photogrammetric Engineering and Remote Sensing, 1990, 56(8): 1160-1169.

[4] Benz U.C., Hofmann P., Willhauck G., Lingenfelder I., Heynen

M.. *Multi-resolution, Object-oriented Fuzzy Analysis of Remote Sensing Data for GIS-ready Information ISPRS*[J]. Journal of Photogrammetry and Remote Sensing, 2004, 58(3): 239-258.

[5] Chanussot J., Benediktsson J.A., Fauvel M.. *Classification of Remote Sensing Images from Urban Areas Using a Fuzzy Possibilistic Model*[J]. IEEE Geoscience and Remote Sensing Letters, 2006, 3(1): 40-44.

[6] Yang C., Bruzzone L., Sun F.Y., Lu L.J., Guan R.C., Liang Y.C.. *A Fuzzy-Statistics-Based Affinity Propagation Technique for Clustering in Multispectral Images*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(6): 2647-2659.

[7] Bruzzone L., Chi M., Marconcini M. *A Novel Transductive SVM for Semisupervised Classification of Remote-sensing Images*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(11): 3363-3373.

[8] Gomez-Chova L., Camps-Valls G., Bruzzone L.. *Mean Map Kernel Methods for Semisupervised Cloud Classification*[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(1): 207-220.

[9] 周志华, 王珏. 机器学习及其应用[M]. 北京：清华大学出版社, 2007.

[10] Miller D.J., Uyar H.S.. *A Mixture of Experts Classifier with Learning based on both Labeled and Unlabelled Data*[M]. Advances in Neural Information Processing Systems 9, Cambridge, MA:MIT Press, 1997: 571-577.

[11] Rosenberg C., Hebert M., Schneiderman H.. *Semi-Supervised Self-training of Object Detection Models*[J]. In Seventh IEEE Workshop on Applications of Computer Vision, 2005, 1: 29-36.

[12] Ando R.K., Zhang T.. *A High-Performance Semi-Supervised Learning Method for Text Chunking*[C]. Proceedings of the 43rd Annual Meeting of the ACL, 2005: 1-9.

[13] McClosky D., Charniak E., Johnson M.. *Effective Self-Training for Parsing*[C]. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, 2006: 152-159.

[14] Maulik U., Chakraborty D.. *A Self-trained Ensemble with Semisupervised SVM: An Application to Pixel Classification of Remote Sensing Imagery*[J]. Pattern Recognition, 2011, 44: 615-623.

[15] Petropoulos G.P., Kalaitzidis C., Vadrevu K.P.. *Support Vector Machines and Object-based Classification for Obtaining Land-use/Cover Cartography from Hyperion Hypspectral Imagery*[J]. Computers & Geosciences, 2012, 41: 99-107.

[16] Jain A.K., Dubes R.C.. *Algorithms for Clustering Data*[M]. Prentice-Hall Advanced Reference Series, 1988.

[17] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[18] Jain A.K., Murty M.N., Flynn P.J.. *Data Clustering: A Review*[R]. ACM Computing Surveys, 1999, 31(3): 264-323.

[19] Jain A.K., Duin R.P.W., Mao J.C.. *Statistical Pattern Recognition:A review*[J]. IEEE Transaction on Pattern Analysis and

Machine Intelligence, 2000, 22(1): 4-37.

[20] Sambasivam S.,Theodosopoulos N.. *Advanced Data Clustering Methods of Mining Web Documents*[J]. Issues in Informing Science and Information Technology, 2006(3): 563-579.

[21] MacQueen J.. *Some Methods for Classification and Analysis of Multivariate Observations*[J].In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1:281-297.

[22] 杨晨. 基于机器学习的土地覆盖遥感信息提取方法研究[R]. 吉林大学, 2010.

[23] Wu J., Lin Z.K., Lu M.U.. *Asymmetric Semi-Supervised Boosting for SVM Active Learning in CBIR*[C]. Proceedings of the ACM International Conference on Image and Video Retrieval. Xian, China, 2010: 182-188.

[24] Sweet J.N.. *The Spectral Similarity Scale and Its Application to the Classification of Hyperspectral Remote Sensing Data*[C]. IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, 2004: 92-99.

[25] 黄金杰, 李士勇, 蔡云泽. 一种建立粗糙数据模型的监督模糊聚类方法[J]. 软件学报, 2005, 16(6): 744-753.

[26] Sun M.C., Chou C.H.. *A Modified Version of the K-means Algorithm with a Distance based on Cluster Symmetry*[J]. IEEE Transaction Pattern Analysis and Machine Intelligence, 2001, 23(6): 674-680.

[27] Gustafson D.E., Kessel W.C.. *Fuzzy Clustering with Fuzzy*

Covariance Matrix[C]. In Proceedings of the IEEE CDC, San Diego, 1979: 761-766.

[28] Bezdek J.C.. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

[29] Xie X.L., Beni G.A.. *Validity Measure for Fuzzy Clustering*[J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 1991, 3(8): 841-847.

[30] Kavzoglu T.. *Determination of Environmental Degradation Due to Urbanization and Industrialization in Gebze, Turkey*[J]. Environmental Engineering Science, 2008, 25: 429-438.

[31] Nigam K., McCallum A., Thrun S.. *Text Classification from Labeled and Unlabeled Documents using EM*[J]. Machine Learning, 1999, 39: 103-134.

第6章

基于半监督集成支持向量机的土地覆盖分类研究

6.1 概述

半监督学习利用未标记样本所隐含的地物类型在特征空间中的结构信息，拟合出一个更有代表性的分类器[1]~[3]；集成学习[4]综合多个同构或异构学习机对同一个问题进行学习，进而提高分类器的泛化能力，二者采用截然不同的思维观，其发展几乎是并行的，只有少数研究涉及二者的结合[5]~[7]。但是，集成学习与半监督学习之间存在许多互补关系[8]~[9]，具体表现在：

(1) 当单个分类器无法通过无标签样本参与训练提高分类精度时，分类器的融合可以进一步改善单个分类器的不足；

(2) 使用未标记数据能够增加个体分类器之间的差异性，从而可获得高效的集成分类模型；

(3) 分类器的集成能够比单个分类器更快地达到理想的分类精度。因此，如何设计有效的半监督集成方案是改进学习系统泛化能力的另一个崭新思路[10]~[12]。

第5章通过自适应半监督学习把大量无标签样本所包含的数据特征加入SVM学习算法的设计中，弥补单个监督分类器的不足，同时也产生若干性能差异的个体分类器。本章探讨如何将这些个体分类器集成，进一步提高分类模型的泛化能力。

6.2 集成学习框架

Krogh 和Vedelsby[13]给出神经网络集成泛化误差计算公式，假

设学习任务是由N个神经网络组成对f：$R^n \rightarrow R$进行近似，集成采用加权平均，各分类器分别赋以权值w_a，并且满足如下条件：

$$\sum_a w_a = 1,\ w_a > 0 \tag{6-1}$$

再假设训练集的每个样例按概率分布$p(x)$随机抽取，分类器a对输入X的输出为$V^a(X)$，则集成输出为：

$$\overline{V}(X) = \sum w_a V^a(X) \tag{6-2}$$

神经网络a的泛化误差E^a和神经网络集成的泛化误差E分别为：

$$E^a = \int p(x)(f(x) - V^a(x))^2 dx \tag{6-3}$$

$$E = \int p(x)(f(x) - \overline{V}(x))^2 dx \tag{6-4}$$

分类器个体a的差异度A^a和集成的差异度$\overline{A}$分别为：

$$A^a = \int p(x)(V(x) - \overline{V}(x))^2 dx \tag{6-5}$$

$$\overline{A} = \sum_a w_a A^a \tag{6-6}$$

则最终的集成泛化误差为：

$$E = \overline{E} - \overline{A} \tag{6-7}$$

$\overline{E}$表示个体分类器固有误差，$\overline{A}$表示个体分类器之间的差异。该公式表明要获得较好的集成结果就需要降低个体分类器的误差并增加个体分类器之间的差异。因此，集成学习框架的设计通常由以下两部分构成：采用一定的个体生成方法，研究如何根据指定训练集生成集成中多个、有差异个体分类器；采用一定的结论合成方法，研究怎样将多个个体分类器的输出进行合成，得到集成学习的最终结果。

6.2.1　个体生成方法

如何构造集成中的个体分类器对集成的性能有重大影响，目前

的方法主要有以下9种。

1. 基于数据划分方法

通过划分训练样本集合产生多个训练样本子集，学习算法分别在这些训练样本子集上进行训练，生成多个个体分类器。典型算法主要有Bagging算法[14]和Boosting算法[15]，它们也是目前集成学习算法中最著名的两种方法。Lima等人[16]将Bagging集成技术引入到SVM回归分析中，实验证明其泛化能力有所提高。He等人[17]比较了不同SVM集成方法，并在Boosting算法基础上给出一种DBoosting算法。实验结果表明，DBoosting算法比其他算法有更好的性能。

2. 基于属性划分方法

把输入特征空间划分为多个特征子集，在不同特征子集上投影得到的训练样本用于训练生成多个个体分类器。李国正等人[18]运用特征选择对Bagging方法中bootstrap方法产生的训练样本子集进行特征选择，分别提出PRIFEB和MIFEB两种算法，从而提高了个体的差异度和精度。Brylia等人[19]提出Attribute-Bagging方法，利用随机方法产生特征子集来获得集成个体之间更大的差异度。

3. 基于分类器模型参数方法

分类参数对分类精度有重要影响，设置不同参数值可以获得不同泛化能力的分类模型。何灵敏[20]提出基于参数与样本二重扰动机制的集成学习算法，在Bagging和Boosting算法中嵌入高斯核函数宽度参数λ和惩罚参数c扰动机制。Li等人[21]将扰动径向基核函数宽度参数方式引入到AdaBoost算法中，利用新的SVM集成算法产生有差异的个体SVM，从中选择出部分差异度较大的个体参与集成，试验结果表明该方法能显著提高分类精度。

6.2.2　结论生成方法

结论生成方法主要研究如何对集成中个体分类器所给出的结论进行合成。目前主要采用的方法有以下几种。

1. 全部生成的分类器个体都参与集成的投票法

主要包括多数投票法和加权投票法，通常投票法更适合分类集成[22]。

(1) 多数投票法(Majority Voting)

基本思想是多个基分类器进行分类预测，通过某种投票原则进行投票表决，是最简单，也是最普遍的结论合成方法。每个成员分类器对于待测样本x有一个类别的判断，并给所判断的待测样本x的归属类别投一票，设N_j为将待测样本x判定给第j类的个体分类器个数，那么最后的判决函数是：

$$f(x)=\arg\max(N_j) \tag{6-8}$$

(2) 加权投票法(Weighted Voting)

加权投票法是将每个成员分类器均赋予一定的权重，权重通过在训练集上测量每个成员分类器精度获得，且权重与精度成正比，即分类能力好的基分类器被赋予较大的权系数，而分类能力相对差的基分类器赋予较小的权系数，集成的结果取决于加权和。设$h_t(t=1，2，\cdots，T)$是第t个成员分类器的决策函数，$w_t(t=1，2，\cdots，T)$是相应的权重，则最后的决策如下所示：

$$f(x)=sign\left(\sum_{i=1}^{T} w_i h_t(x)\right) \tag{6-9}$$

Yan等人[23]提出类别数目不均的SVM集成，其集成方案是将大样本的负类样本分成K等分，与小样本的正类样本合成一个训练

样本集，在其上训练生成K个体分类器，最后用多数投票法进行集成。Kim等人[24]结合Bagging和Boosting生成个体SVM分类器，并比较了多数投票法、加权投票法和两层SVM3种分类方式，实验结果表明，3种方法都能获得比单个SVM更好的泛化性能。

2. 从集成中选择出部分个体的选择性集成学习算法

选择集成算法基于“Many Could Be Better Than All”[25]，基本思想就是利用对多个学习器进行适当的选择来剔除对学习系统有副作用的学习器，最后将所选择的结果进行结合，从而得到性能更好的学习器。Zhou等人[26]提出了相应的选择性集成算法GASEN，其理论分析和实验结果均表明，该算法性能优于Boosting和Bagging算法。He等人[27]提出基于遗传算法的选择性SVM集成方法，从集成中选择出部分个体参与集成，得到比全部个体都参与集成更好的性能。

6.3 半监督集成支持向量机的土地覆盖分类模型构建

针对上述分析，本文从个体生成(使用程序来生成个体分类器)和结论生成(选择特定的策略来组合分类器)两个部分考虑，提出半监督集成SVM分类策略。具体技术路线如下：第一，个体生成部分一方面利用自适应变异粒子群算法(Self-adaptive Mutation PSO，SAMPSO)优化SVM分类器参数以获得高精度分类器个体，另一方面采用Gustafson-Kessel(GKclust)模糊聚类算法控制Self-training错误标记样本的加入以提高个体分类器的差异性；第二，结论生成部

分采用加权投票策略将半监督分类器个体集成。具体描述如图6-1所示：

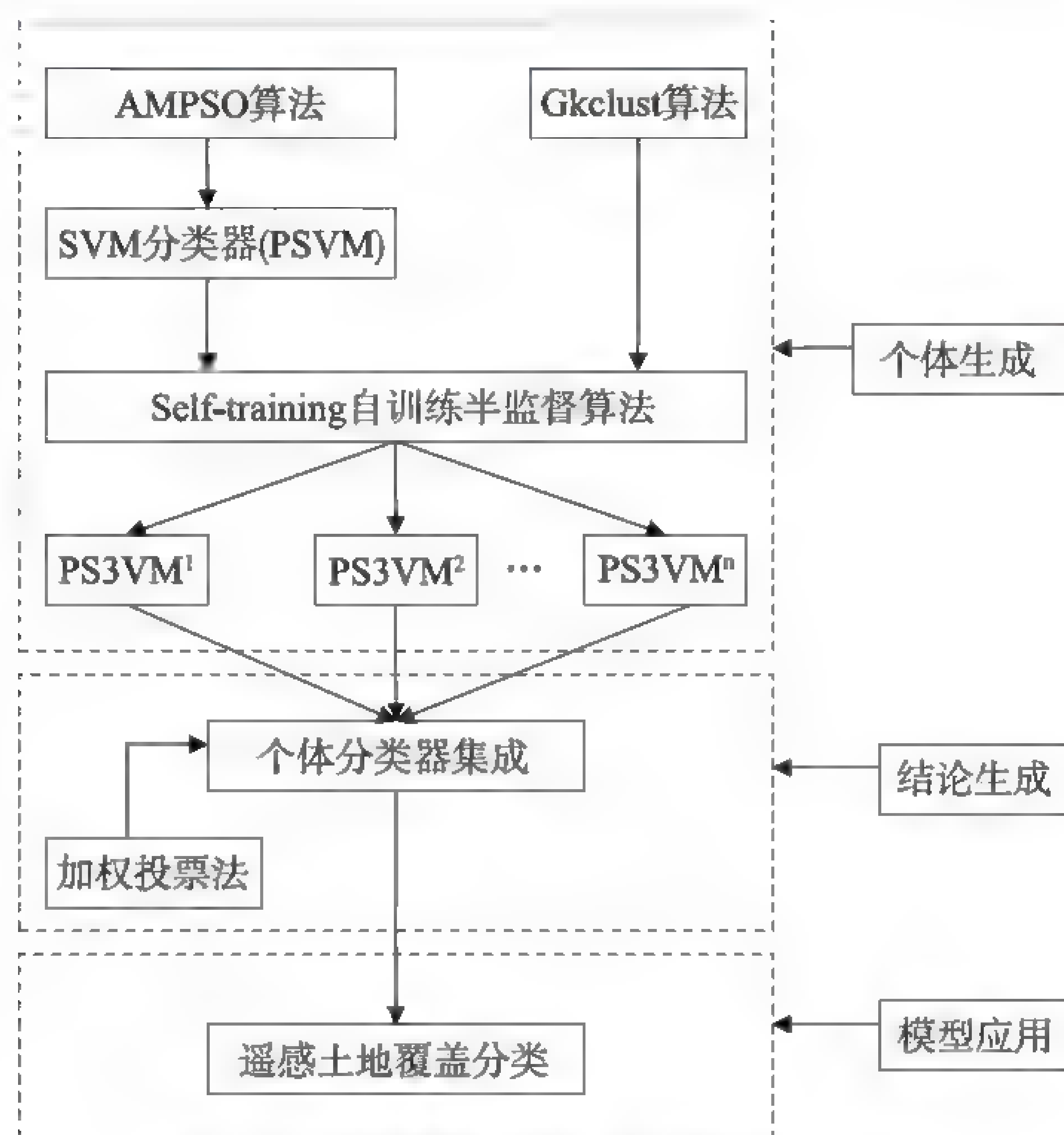


图6-1　半监督集成SVM分类模型技术路线

6.3.1　个体生成算法

第5章所构建的PS3VM分类模型，以PSVM作为自训练算法的基分类器，在自训练算法未标记样本标注过程中，利用GKclust模糊聚类算法控制错误类别样本的标注，提高分类器的分类精度，同时也迭代产生了多个半监督分类器个体$PS3VM^1$，$PS3VM^2$，…，$PS3VM^n$(算法详见第5章)。

6.3.2 结论生成算法

集成学习主要思想是利用分类器的融合改善单个分类器的不足，其性能一方面取决于多样性强的个体学习器，另一方面依赖于成员分类器的有效组合。接下来专著利用加权投票法将这些个体分类器集成，进一步提高分类模型的泛化能力。在此，将集成的半监督分类模型简称EPS3VM。

假设利用自训练算法产生T个分类器个体：S^1，S^2，…，S^T；一幅遥感影像分类问题包含C个类别。

EPS3VM算法步骤如下：

Step1：将各基分类器S^1，S^2，…，S^T的分类混淆矩阵获得的各类别的用户精度作为权值W^i_j(i=1，2，…，T；j=1，2，…，C)；

Setp2：各基分类器对未知像元X分类后，将分类结果相同的各基分类器对该类别的权值相加，即得到像元X属于各类别的权值之和$\sum_{i=1}^{T} W_j^i (j=1, 2, \cdots, H)$；

Setp3：比较权值之和的大小，将最大值对应的类别作为像元X的最终类别标签。

6.4 实验结果与分析

为了测试EPS3VM分类模型性能，将半监督集成模型应用于多光谱遥感影像的土地覆盖分类实验，同时与PSVM、PS3VM进行对比。

6.4.1　实验数据

本文选择2006年9月22日获取行列号115-30多光谱Landsat-5 TM遥感影像(30米空间分辨率，UTM投影)。根据植被的光谱特征和空间分布规律，本文提取了8个特征，包括TM图像的6个波段(1～5，7)、PCA变换的第一主分量、植被指数(NDVI)。

据研究区实际情况并参照第3章介绍的土地覆盖分类系统一级类型，将实验区分为5个土地覆盖类型，即森林、水体、农田、人工表面及其他。为了保证每个类别数据的变化性和代表性，数字集采用随机像素的选择策略。土地覆盖类型及样本数量如表6-1所示。

表6-1　类别及样本数量

类别代号	类别名称	样　本
ω_1	森林	334
ω_2	水体	229
ω_3	农田	268
ω_4	人工表面	190
ω_5	其他	229
类别及样本总数	5	1 250

6.4.2　结果与精度分析

当训练样本较少时，未标记样本参与的半监督分类方法可有效提高分类精度，但随着已标记分类样本的增加，未标记样本的作用越来越小。首先为了更好地体现小样本特点，将随机抽取少部分样本作为训练样本(占每类样本的30%)，整个数据集用于测试。分别采用PSVM、PS3VM、EPS3VM 3种分类模型进行对比实验，将分类精度、Kappa系数及相应的参数值列于表6-2。实验结果表明，使用EPS3VM方法分类得到的分类精度比PS3VM模型高出4.72%，

比PSVM模型高出8.4%，Kappa系数也要分别高于PS3VM模型0.059 6，高于PSVM模型0.106。表6-3显示3种方法不同类别的混淆矩阵，实验结果均证实EPS3VM能有效提高影像的分类精度。

表6-2　分类参数、分类精度和kappa系数的比较

分类模型	惩罚参数c	核函数参数λ	分类精度(%)	Kappa系数
PSVM	52.631 2	27.103 1	88.48	0.854 6
PS3VM	24.060 2	12.892 3	92.16	0.901 0
EPS3VM	35.632 2	28.104 3	96.88	0.960 6

表6-3　三种方法的混淆矩阵

PSVM	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	334	0	0	0	0
ω_2	0	169	0	0	60
ω_3	5	2	238	3	20
ω_4	0	0	15	166	9
ω_5	3	0	0	27	199
PS3VM	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	334	0	0	0	0
ω_2	0	194	0	0	35
ω_3	3	2	251	4	8
ω_4	0	0	12	170	8
ω_5	5	0	0	21	203
EPS3VM	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	334	0	0	0	0
ω_2	0	229	0	0	0
ω_3	1	0	263	1	3
ω_4	0	0	12	172	6
ω_5	3	0	0	13	213

上述实验利用遥感影像的数字集，有效地测试了EPS3VM的性能。在本实验中，将3种分类模型应用于覆盖研究区域115-30的TM影像子集的分类，并产生分类专题图，分类结果如图6-2所示。由于研究区主要土地覆盖类型为植被，因此，图6-2(a)为研究区5、4、3

波段合成图；图6-2(b)显示PSVM分类结果，其中1 250个样本点作为训练集；图6-2(c)和图6-2(d)以1 250样本点作为已知样本点，并随机从影像上搜集3 000个未知标签点进行半监督分类和半监督集成分类。从分类图中可以看出，研究区域的主要土地覆盖类型为林地。比较3种分类图，不难发现PSVM分类模型在林地和耕地存在错分现象，PS3VM主要问题在于耕地、裸地和住宅分类错误，而EPS3VM在遥感影像数据上的分类明显优于其他两种方法。

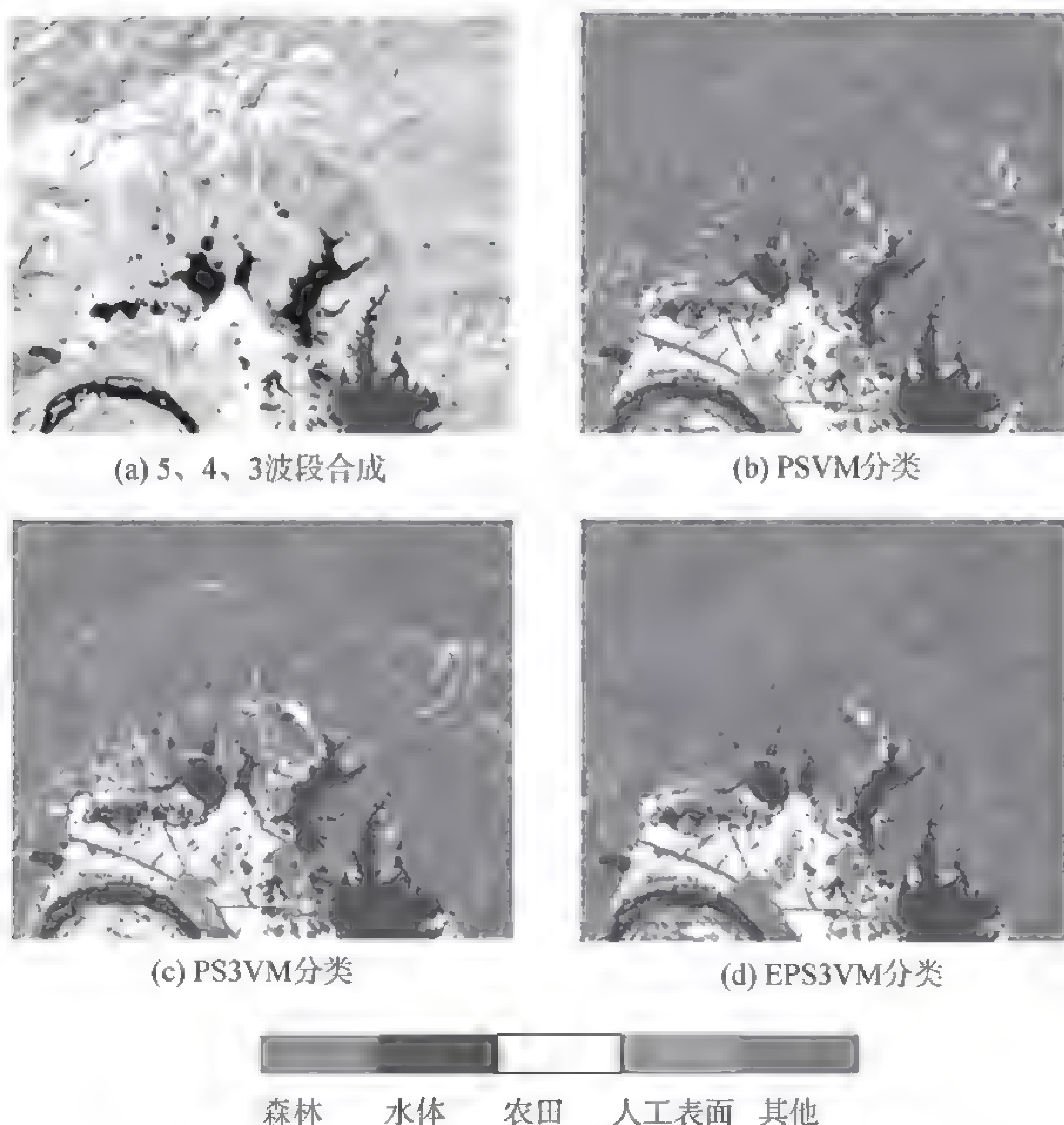


图6-2　三种分类方法对TM影像子集产生的分类专题图

6.5　本章小结

构建性能良好的分类器集成，可以获得比单个分类器更好的学习性能和泛化能力。通常需要满足的充要条件是：

(1) 需要足够高的个体分类器精度，其分类精度是指对于一个新的数据进行函数逼近或分类，它的误差率要比随机猜测好；

(2) 个体分类器之间要有一定差异度，其差异度指的是对于新的数据点进行函数逼近或分类时，它们的错误是不相关的。

本书从以上两个方面考虑，提出基于半监督集成SVM分类方法。该方法有如下特点：

(1) 以PSVM模型作为基分类器，利用Self-training算法产生半监督分类器个体，其中在半监督学习过程中为了避免错误样本的加入引入GKclust模糊聚类算法。

(2) 半监督学习和集成学习两种范式的结合，一方面充分利用大量廉价的未标记样本，以减少对有标记样本的需求量；另一方面，未标记数据能够增加个体分类器之间的差异性，从而进一步提高学习系统的泛化能力。

(3) 利用所提模型解决遥感土地覆盖分类实验表明，在相同样本数量条件下，相比于其他分类技术，该模型能获取更丰富、更准确的遥感类别信息。同时，本文算法亦存在一些有待改进之处。例如，在个体分类器集成部分，对权系数产生的问题上，可采用更有效的策略。

参考文献

[1] Chapelle O., Schölkopf B., Zien A.. *Semi-Supervised Learning*[M]. MIT Press, Cambridge, MA, 2006.

[2] Zhou Z.H., Li M.. *Semi-supervised Learning by Disagreement*[J]. Knowledge and Information Systems, 2010, 24(3): 415-439.

[3] Zhu X.. *Semi-supervised Learning Literature Survey*[R]. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

[4] Zhou Z.H.. *Ensemble Learning*[M]. Encyclopedia of Biometrics, Springer, Berlin, 2009: 270-273.

[5] 邬俊, 段晶, 鲁明羽. 基于偏袒性半监督集成的SVM主动反馈方案[J]. 模式识别与人工智能, 2010, 23(6): 745-751.

[6] Mallapragada P.K.R., Jin R., Jain A.K., Liu Y.. *SemiBoost: Boosting for Semi-supervised Learning*[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2000-2014.

[7] Valizadegan H., Jin R., Jain A.K.. *Semi-supervised Boosting for Multi-class Classification*[C]. In Proceedings of the 19th European Conference on Machine Learning, 2008: 522-537.

[8] Bennett K., Demiriz A., Maclin R.. *Exploiting Unlabeled Data in Ensemble Methods*[C]. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002: 289-296.

[9] Zhou Z.H.. *When Semi-Supervised Learning Meets Ensemble*

Learning[J]. Frontiers of Electrical and Electronic Engineering in China, 2010, 6(1): 6-16.

[10] Dalché-Buc F., Grandvalet Y., Ambroise C.. *Semi-supervised MarginBoost*[J]. Advances in Neural Information Processing Systems, 2002, 14: 553-560.

[11] Wu J., Lin Z.K., Lu M.U. *Asymmetric Semi-Supervised Boosting for SVM Active Learning in CBIR*[C]. Proceedings of the ACM International Conference on Image and Video Retrieval. Xian, China, 2010, 182-188.

[12] Zhou Z.H., Li M.. *Semi-supervised Learning with Co-training*[C]. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Scotland, 2005: 908-913.

[13] Krogh A., Vedelsby J.. *Neural Network Ensembles, Cross Validation, and Active Learning*[J]. Advances in Neural Information Processing Systems, 1995, 7: 231-238.

[14] Breiman L.. *Bagging Predictors*[J]. Machine Learning, 1996, 24(2): 123-140.

[15] Freund Y., Sehapire R.E.. *A Decision-theoretic Generalization of On-line Learning and an Application to Boosting*[J]. Journal of Computer and System Sciences, 1997,55: 119-139.

[16] Lima C.A.M., Coelho A.L.V., Von Zuben E.J.. *Ensembles of Support Vector Machines for Regression Problems*[J]. INNS-IEEE International Joint Conference Oil Neural Networks, 2002, 3: 2381-2386.

[17] He L.M., Yang X.B., Kong F.S.. *Support Vector Machines*

Ensemble with Optimizing Weights by Genetic Algorithm[C]. Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, 2006: 3503-3507.

[18] 李国正, 李丹. 集成学习中特征选择技术[J]. 上海大学学报(自然科学版), 2007, 13(5): 598-604.

[19] Brylia R., Gutierrez Osunab R., Queka F.. *Attribute Bagging: Improving Accuracy of Classifier Ensembles by Using Random Feature Subsets*[J]. Pattern Recognition, 2003, 36 (6): 1291-1302.

[20] 何灵敏. 支持向量机集成及其在遥感分类中的应用[J]. 浙江大学, 2006.

[21] Li X.C., Wang L., Sung E.. *AdaBoost with SVM Based Component Classifiers*[J]. Engineering Application of Artificial Intelligence, 2008, 21: 785-795.

[22] 杨静宇. 分类器集成研究[R]. 南京理工大学, 2009.

[23] Yan R., Liu Y., Jin R.. *On Predicting Rare Classed with SVM Ensembles in Scene Classification*[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: 6-10.

[24] Kim H., Pang S., Je H.. *Constructing Support Vecotr Machine Ensemble*[J]. Pattern Recognition, 2003, 36(12): 2757-2767.

[25] Zhou Z.H., Wu J.X., Tang W.. *Ensembling Neural Networks: Many Could Be Better Than All*[J]. Artificial Intelligence, 2002, 137(1-2): 239-263.

[26] Zhou Z.H., Tang W., Chen Z.Q.. *Combining Regression Estimators: GA-based Selective Neural Network Ensemble*[J]. International Journal

of Computational Intelligence and Applications, 2001, 1(4): 341-356.

[27] He L.M., Yang X.B, Lu H.J.. *A Comparison of Support Vector Machines Ensemble for Classification*[C]. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, 2007: 3613-3617.

第7章

总结与展望

7.1 研究结论

土地覆盖变化的研究是全球环境研究的热点，与人类生产生活、生态环境密切相关。土地覆盖遥感信息的获取不仅要靠日益发展的遥感技术，更要依赖于先进的信息提取方法。因此，探索新的遥感分类理论与分类方法一直是国内外学者关注的焦点。本书以现有SVM为基础，深入分析其遥感分类领域的研究现状及不足，融合智能算法、模糊聚类算法、半监督学习、集成学习理论，探讨新的土地覆盖遥感分类方法。实验证明，智能算法、模糊理论、半监督学习、集成学习理论的引入，实现了算法的相互融合，优势互补，进一步提高了土地覆盖遥感信息提取的精度，丰富了机器学习理论在遥感信息提取与识别方面的理论基础和应用研究。对于土地覆盖遥感数据信息提取具有较强的理论意义和实用价值。

本书的主要研究内容及结论如下：

(1) 参数选择是任何机器学习方法都必须面对的重要而困难的问题。设置合适的SVM参数对提高遥感影像分类精度是非常重要的。针对这一问题，专著提出一种自适应变异粒子群(Adaptive Mutation Particle Swarm Optimization，AMPSO)优化SVM参数模型。与传统粒子群算法(Particle Swarm Optimization，PSO)相比，AMPSO在运行过程中根据群体适应度方差以及当前最优解的大小来确定当前最佳粒子的变异概率，避免了传统PSO因早熟收敛造成分类参数寻找不准确的缺点。最后应用该模型进行多光谱遥感影像的土地覆盖分类试验，并与最大似然法、SVM分类方法、传统PSO-SVM 3种分类方法进行对比实验，结果表明所提模型能够有效提高

遥感影像的分类精度。

(2) 准确且充足的训练样本是影响SVM分类精度的另一个重要因素。遥感影像本身具有高度的复杂性和随机性特点，此外，训练样本是分析者人为选取的，不论现场踏勘还是参考已有专题图件或其他资料，总是存在经验、知识有限和盲目选择的缺点，从而造成所选择的分类样本不足且无法保证对影像中相应地物的代表性。针对这一问题，本书提出了一个新的自训练PSVM方法，利用自训练半监督技术充分挖掘未标记样本所蕴涵的结构信息，以此对已标记样本数量不足且代表性不好而造成的拟合分类器有偏差的情况进行矫正。为了解决遥感影像混合像元分类困难问题，在标注未标记样本阶段，引入模糊聚类技术，以控制错误信息的输入，同时利用自适应变异粒子群算法对SVM参数优化以提高半监督模型基分类器的分类精度。为了测试提出模型的有效性，分别针对遥感的数字集和影像集进行分类实验，并与传统SVM监督分类方法、未改进自训练半监督SVM方法进行对比实验，实验结果一方面说明了已标记样本和未标记样本的用量比例必须满足一定的阈值要求，才能产生最小的泛化误差；另一方面证实了利用所提出学习框架能够有效地克服由于样本不足且代表性不好而导致分类精度低下的问题。

(3) 为了进一步提高分类器的泛化能力，本书将集成学习技术引入半监督分类模型，提出了一种新的半监督集成方案，主要设计思想在于：半监督分类方法利用未标记数据有效地应对训练样本不足缺点的同时，也产生若干性能差异的个体分类器，将这些个体分类器加权投票集成。为了测试其性能，应用该模型进行多光谱遥感影像的土地覆盖分类实验，结果表明，半监督技术与集成技术的有

效融合能进一步改善分类器的分类精度。

综上所述，本书将智能算法、模糊集理论以及半监督与集成学习引入SVM分类中，提高了SVM分类器的分类精度，丰富了土地覆盖信息提取方法，促进了相关领域的发展和进步。

7.2 本书不足之处

本书利用半监督学习及集成学习理论融合SVM分类器，解决了SVM分类模型在遥感影像分类中所表现的不足，对其土地覆盖分类提取技术研究有一定意义，然而，仍存在如下有待改进之处：

(1) 遥感影像的数字化部分，考虑到时间复杂度，所选分类特征略显不足；遥感数据源单一，仅针对一景的多光谱遥感影像，缺乏多数据源的融合。

(2) 所提理论在遥感的分类精度上有所提高，然而时间复杂度却不是很理想，分类模型时间复杂度有待进一步提高。

7.3 研究展望

本书对半监督集成SVM方法在土地遥感分类研究领域做了初步的探讨，形成了初步的成果。鉴于该理论的优点，未来必将促进土地遥感分类技术及相关领域的进步和发展。

(1) 半监督集成SVM模型能有效应对小样本问题，适合解决如

跨境地区的影像的识别，因为跨境区域的样本获取条件更为苛刻，无法现场获取已知样本点，境外样本的来源便寄托于与境内影像条件相似的覆盖区域的影像，由此造成样本数量不足且代表不好。

(2) 多种学习技术融合的思想进一步扩宽算法研究的思路。随着计算机技术的不断发展，遥感分类方法层出不穷，每个分类算法都各具优缺点。将有效的算法相互融合，取长补短，更具有实际的应用价值。